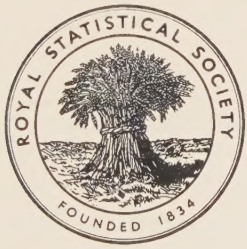
APPLIED STATISTICS

A Journal of the Royal Statistical Society

3043 PERIODICAL
Applied Statistics
V. 9-10
1960-61

EDITOR
DONALD G. BEECH

ASSISTANT EDITOR
HAROLD S. BOOKER



VOLUME NINE

Oliver and Boyd Ltd.
London: 39A Welbeck Street, W.1
Edinburgh: Tweeddale Court

1960

Digitized by the Internet Archive
in 2025

https://archive.org/details/journal-of-the-royal-statistical-society-series-c_1960_9

CONTENTS

No. 1. MARCH

Page

Some Applications of Multivariate Methods in Botany
S. Clifford Pearce and Derek A. Holland 1

Some Statistical Aspects of Preference and Related Tests
C. I. Bliss 8

A Study of Some Potential Biases in the Operation of a Consumer Panel .. .. *A. S. C. Ehrenberg* 20

An Application of Linear Programming to Agricultural Economics
Shirley E. Vincent 28

On Comparing Two Observed Frequency Counts *Morton Kupperman* 37

Analysis of Variance of a Latin Square Design with Missing Observations .. .. *Clyde Young Kramer and Suzanne Glass* 43

Note on British Unemployment Statistics .. .. *Donald Bailey* 51

Meetings of Sections and Local Groups of the Royal Statistical Society 60

Book Reviews:

Economics for Pleasure, G. L. S. Shackle 62

International Journal of Abstracts—Statistical Theory and Method, W. R. Buckland (General Editor) 63

Statistical Methods in Biology, Norman T. J. Bailey 64

Operations Research: An Annotated Bibliography, James H. Batchelor 66

Modern Business Statistics, John E. Freund and Frank J. Williams 67

Sampling Inspection Tables, Single and Double Sampling
Harold F. Dodge and Harry G. Romig 67

Publication Received 68

No. 2. JUNE

The Economic Incentive Provided by Sampling Inspection
I. D. Hill 69

Some Problems in Applying Statistics to Geology
W. C. Krumbein 82

The Consistency of Setting of a Machine Tool Handwheel
E. Nigel Corlett and Geoffrey Gregory 92

Statistics and Taste Testing .. .. *Norman T. Gridgeman* 103

The Economics of New Roads with Particular Reference to M1
T. K. Burdess 113

Interviewer Variability and a Budget Survey .. *W. F. F. Kemsley* 122
Notes and Comments 129

Book Reviews:

Analysis of Straight-Line Data, Forman S. Acton 130
Regression Analysis, E. J. Williams 130
Publication Received 132

No. 3. NOVEMBER

Stock Control Problems: Some Fallacies in their Current Treatment
Eric Duckworth 133
Classification Procedures Based on Bayes's Formula
A. Birnbaum and A. E. Maxwell 152
Operational Statistics in Instrumental Analysis
Allan B. Calder 170
Physical Inventory using Sampling Methods .. *Marion R. Bryson* 178
Analysis of Variance of a Balanced Incomplete Block Design with Missing Observations *Hugh Robert Baird and Clyde Young Kramer* 189
Meetings of Sections and Local Groups of the Royal Statistical Society 199

Book Reviews:

The Statistical Basis of Quality Control Charts, S. K. Ekambaram 201
Introduction to Quantitative Genetics, D. S. Falconer 202
Rapid Statistical Calculations, M. H. Quenouille 203
Tables of Random Normal Deviates
J. M. Sengupta and Nikhilesh Bhattacharya 204

SOME APPLICATIONS OF MULTIVARIATE METHODS IN BOTANY

S. CLIFFORD PEARCE and DEREK A. HOLLAND
East Malling Research Station, Kent

The techniques of component analysis and factor analysis are here reviewed in relation to biological problems, and their utility in this context is studied by considering their application to some of the problems of fruit tree growth and cropping.

Over the last few years multivariate methods have come into prominence, and for the first time there have been the computing facilities to make extensive use of them. Although the techniques of factor analysis and component analysis have, in the past, generally been associated with psychological applications, there are many other fields involving the simultaneous study of a number of variates in which they could be of assistance—a suggestion already made in this journal.[3] This is certainly true of the biological field. Here, for example, an experimenter may have values for the concentrations of a number of elements in a tissue, and as a result of a series of univariate analyses a statistician may be able to tell him that a certain treatment has 'significantly increased the nitrogen concentration but lowered the potassium, at the same time increasing (but not significantly) the phosphate', and so on. This alone does not give him much insight into the general position. Several things could have happened to the potassium: for example, there could have been (i) a general decrease in cations, or (ii) a substitution of calcium for potassium, or (iii) an increase in growth with the total potassium content remaining constant, bringing about a dilution of potassium. To the experimenter the interest must lie in the relationships between the variates, and little is learnt of these by a series of univariate analyses.

As an attempt to overcome this difficulty we have considered the application of both component and factor analysis to some of the problems associated with the growth and cropping of fruit trees. Standard methods of analysis have been employed[4] and no attempt has been made to extend them. The aim has been a critical study of their usefulness, and the purpose of this paper is to set out some of the results obtained.

Component Analysis

The simpler method is component analysis; thereby factors are suggested that could have given rise to the dispersion matrix, i.e. that of variances and covariances, or the correlation matrix. However, before deciding that the factors suggested are genuine, it is necessary

to consider if (i) they are found consistently and (ii) they are biologically reasonable. At this stage close liaison is needed between statistician and biologist, or absurd conclusions will be reached and the approach brought into disrepute.

In a recently published study[7] of the four quantities:

w = log (weight of mature tree above ground)
g = log (basal trunk girth of mature tree)
l = log (total shoot growth during the first four years)
e = log (basal trunk girth at four years of age)

for two varieties of apples the correlation matrices proved to be:

$$\begin{bmatrix} 1{\cdot}000 & 0{\cdot}951 & 0{\cdot}596 & 0{\cdot}517 \\ 0{\cdot}951 & 1{\cdot}000 & 0{\cdot}694 & 0{\cdot}619 \\ 0{\cdot}596 & 0{\cdot}694 & 1{\cdot}000 & 0{\cdot}898 \\ 0{\cdot}517 & 0{\cdot}619 & 0{\cdot}898 & 1{\cdot}000 \end{bmatrix}$$

and

$$\begin{bmatrix} 1{\cdot}000 & 0{\cdot}939 & 0{\cdot}266 & 0{\cdot}178 \\ 0{\cdot}939 & 1{\cdot}000 & 0{\cdot}424 & 0{\cdot}358 \\ 0{\cdot}266 & 0{\cdot}424 & 1{\cdot}000 & 0{\cdot}835 \\ 0{\cdot}178 & 0{\cdot}358 & 0{\cdot}835 & 1{\cdot}000 \end{bmatrix}$$

These are apparently different. However, if component analysis is used, the first may be accounted for by the operation of two factors with the following vectors:

$$\zeta_1 = 0{\cdot}489\ w + 0{\cdot}521\ g + 0{\cdot}507\ l + 0{\cdot}482\ e$$
$$\zeta_2 = 0{\cdot}564\ w + 0{\cdot}416\ g - 0{\cdot}439\ l - 0{\cdot}562\ e$$

which explain respectively 79 and 18 per cent of the variation. The second matrix similarly gives the vectors:

$$\zeta_1 = 0{\cdot}485\ w + 0{\cdot}551\ g + 0{\cdot}497\ l + 0{\cdot}463\ e$$
$$\zeta_2 = 0{\cdot}550\ w + 0{\cdot}409\ g - 0{\cdot}481\ l - 0{\cdot}547\ e$$

which explain respectively 63 and 32 per cent of the variation. Clearly the criterion of consistency has been satisfied with respect to the vectors. The difference in relative importance of the two vectors in the two examples could account for the difference between the two matrices. It is now necessary to ask if the results are meaningful biologically.

For each matrix the first vector gives approximately the same weight to each of the four variates, i.e. it indicates a factor that affects all four growth measurements to about the same degree. In fact some trees grow larger than others however measured. The second vectors equally represent a balance between w and g, which were measured when the trees were mature, and l and e, which were measured when they were only four years old. The factor here is plainly one of establishment, some trees exceeding their early promise and some falling short of it. Since little variation, only three per cent and five per cent

respectively, remains to be explained, there can be no other important source of variation in the four variates.

Sometimes, however, no consistency can be found, and this indicates that different factors operate in the different instances, or perhaps that the variation is not the result of underlying factors at all, but is only the outcome of errors of measuring and sampling. Another possibility is that a set of vectors will appear repeatedly though no reasons can be assigned for its doing so; in such a case the result might have to be accepted as an empirical fact.

If the aim is to discern what are the constant factors in natural phenomena then the attainment of constancy appears to us to be the criterion upon which to decide many of the questions disputed among users of component analysis—e.g. whether to work with the dispersion matrix or the correlation matrix, whether to transform the variates or not. That is not to say that one course may not be more logical than another. Thus, in dealing with size records of plants over a period of time any factors supposedly affect not size itself but growth rates, and a logarithmic transformation appears to be called for. Similarly, a logarithmic transformation may be required to make the relationships linear before the computation of correlation coefficients.

Again, the correlation matrix is really the dispersion matrix for standardised variates, and the question whether to standardise or not may find its answer in the role of the differing standard errors. Are they merely the result of differing scales of measurements or are they part of the problem? If the dispersion matrix is used greater weight will be given to the variates with the higher standard errors and these might be the less interesting ones.

A More Extended Application of Component Analysis

It is well known by horticulturists that trees may vary in a number of characteristics such as size, general shape, habit of growth, etc. In consequence it has been the practice among pomological research workers to measure a large number of different features—trunk girth at various heights, total annual shoot growth, weight of tree where possible, and so on—in order to describe a tree fully. It is also known that many of these features are interrelated one with the other, e.g. trunk girth and tree weight,[5] branch girth and shoot growth.[2] What is really required is to ascertain just what the basic characteristics are (general experience suggests them to be few in number) and what functions of which features can best be used as measures of them. This, of course, is the purpose of component analysis, which can be regarded as being no more than a technique for doing, algebraically and in the general case of n dimensions, what is done, geometrically and in two dimensions, when axes are fitted to an ellipse of points forming a scatter diagram. It would appear that the technique, so regarded, should give, at least, a useful lead into the problem and some insight into the mechanism of tree growth and form.

The measured features in this investigation, which was divided into three stages, were w and g (as in the previous example), k, the logarithm of the upper trunk girth, and q, which equals log (sum of the branch girths/the square root of their number). The bracketed quantity is an approximation to the circumference of a cylinder having the same area of cross-section as that of all the branches together and has been found to be correlated with the basal trunk girth[6] and with the weight of the head of the tree.[1]

Stage 1. Using data from a trial of different apple rootstocks, the correlation matrix based on 'between rootstock' variation gave the vectors:

$$\begin{aligned}\zeta_1 &= 0{\cdot}565\ w + 0{\cdot}403\ g + 0{\cdot}553\ k + 0{\cdot}461\ q\\ \zeta_2 &= 0{\cdot}306\ w + 0{\cdot}688\ g - 0{\cdot}353\ k - 0{\cdot}555\ q\\ \zeta_3 &= -0{\cdot}421\ w + 0{\cdot}472\ g - 0{\cdot}443\ k + 0{\cdot}635\ q\end{aligned}$$

accounting respectively for 60, 27, and 9 per cent of the variation.

The first of these appears to correspond to a general size characteristic but still leaves a considerable part of the variation unaccounted for. Now both the second and third vectors show a contrast of w with q and of g with k. This latter contrast is a simple measure of taper and as such is a characteristic which has been noted before between rootstocks.[8] The contrast of w with q, on the other hand, can be regarded as a balance between total tree weight and a measure of the size of the head of the tree; a characteristic such as would be affected by the removal of large branches as a result of damage, breakage by excessive crops, or measures to keep alleyways clear for the passage of tractors and other farm implements, all of which causes are likely to be associated largely with type of rootstock and the resultant habit of growth. These two characteristics, let them be called Taper (T) and Habit (H), will not be independent, the spreading habit usually being associated with the narrower taper, and therefore could not appear as two orthogonal vectors as the result of a component analysis. Instead it is reasonable to expect two vectors of the form $(T+H)$ and $(T-H)$, as for the second and third vectors observed in this example.

Stage 2. The same data were used again but this time the correlation matrix was based on 'within rootstock' variation, so eliminating the rootstock differences. This gave the vectors:

$$\begin{aligned}\zeta_1 &= 0{\cdot}483\ w + 0{\cdot}524\ g + 0{\cdot}520\ k + 0{\cdot}470\ q\\ \zeta_2 &= -0{\cdot}568\ w - 0{\cdot}092\ g - 0{\cdot}106\ k + 0{\cdot}811\ q\end{aligned}$$

The first of these, again a general size characteristic, now accounted for 86 per cent of the variation. The second, accounting for a further 8 per cent, is a balance between two variates only, the other coefficients being relatively small. The important variates are w and q and thus the vector represents the habit characteristic encountered in the previous stage. The reason for its appearance in this instance, involving only 'within rootstock' variation, is probably associated with

the age of the trees, 33 years. Over such a period there must have been a certain amount of branch removal or breakage distributed in a random fashion, so disturbing the relationship between the weight of the tree and the size of its head.

Stage 3. The same four variates were measured on a set of six-year-old apple trees grown for biometrical investigations. Every effort was made during growth to reduce the external sources of variation.

Nearly all the variation (97 per cent) was accounted for by the single vector:

$$\zeta_1 = 0{\cdot}500\ w + 0{\cdot}500\ g + 0{\cdot}501\ k + 0{\cdot}498\ q$$

so any variation in characteristics other than general size can, in this instance, be regarded as negligible. This is hardly surprising when variation has been controlled as far as possible and the only uncontrollable sources are soil fertility, exposure to the elements, proneness to accident, damage, and disease, etc., factors which might well affect tree size but which are not likely to affect tree shape to any marked degree over such a short period in the life of the tree.

Thus, the method of component analysis leads to the not unreasonable conclusions that young trees grown with a view to uniformity vary only in general size, and that older trees may show also a certain amount of variation in the ratio of the size of head to total weight. Between rootstocks there will, in addition to variation in overall size, be variation in both habit of growth and in taper, i.e. the balance between apical and cambial meristematic activity.

Factor Analysis

In component analysis it is assumed that the variation observed is mainly due to factors having characteristic effects on each of the variates, but this assumption will not always be sound. In factor analysis the possibility of extraneous variation is taken into account, and the series of 'ones' down the leading diagonal of the correlation matrix is replaced by a series of smaller numbers, called 'communalities'; these represent that proportion of the variation of each variate which is due to factors operating on some or all of the other variates, the rest, it is assumed, being due to chance. The criteria that have been advanced for component analysis, namely consistency and intelligibility, apply here with even greater force, because the fitting of the appropriate mathematical model is far more speculative.

An example may be found in the study of successive crops from a set of trees. For example, component analysis of the crops of a set of Lane's Prince Albert apple trees for the years 1939 to 1942 led to only one factor being found, and this accounted for 99·8 per cent of the total variation in the dispersion matrix. Its vector was:

$$\zeta_1 = (0{\cdot}691,\ 0{\cdot}267,\ 0{\cdot}044,\ 0{\cdot}671)$$

while the mean annual crops were in the ratios of:

$$0{\cdot}690 : 0{\cdot}267 : 0{\cdot}046 : 0{\cdot}671$$

Plainly this is a general cropping effect with no complications.

If, however, the same trees were studied for the period 1948 to 1953, component analysis of the correlation matrix in this instance gave three factors with the vectors:

$$\zeta_1 = (0{\cdot}203, 0{\cdot}493, 0{\cdot}268, 0{\cdot}448, 0{\cdot}442, 0{\cdot}497)$$
$$\zeta_2 = (0{\cdot}715, -0{\cdot}273, 0{\cdot}136, -0{\cdot}369, 0{\cdot}475, -0{\cdot}183)$$
$$\zeta_3 = (-0{\cdot}371, -0{\cdot}054, 0{\cdot}883, -0{\cdot}270, 0{\cdot}041, -0{\cdot}065)$$

The first of these represents the general factor of cropping intensity, and the second could well represent a biennial effect, but these account for only 80 per cent of the variation, and the remaining vector, accounting for a further 16 per cent, makes no obvious sense. If a similar analysis is made of the crops of some nearby Conference pear trees over the same period, the results are even more confusing, the first four factors having the vectors:

$$\zeta_1 = (0{\cdot}371, 0{\cdot}421, 0{\cdot}398, 0{\cdot}390, 0{\cdot}462, 0{\cdot}401)$$
$$\zeta_2 = (0{\cdot}533, -0{\cdot}437, -0{\cdot}309, 0{\cdot}544, -0{\cdot}341, 0{\cdot}137)$$
$$\zeta_3 = (-0{\cdot}414, -0{\cdot}277, 0{\cdot}379, 0{\cdot}004, -0{\cdot}350, 0{\cdot}697)$$
$$\zeta_4 = (0{\cdot}147, -0{\cdot}398, 0{\cdot}753, 0{\cdot}068, -0{\cdot}031, -0{\cdot}496)$$

which account for 58, 19, 10, and 7 per cent of the variation respectively. Here the first vector is clear enough, but the interpretation of the others is obscure.

If now 'communalities' are fitted, these prove to be:

$$0{\cdot}80, 1{\cdot}00, 0{\cdot}94, 0{\cdot}75, 0{\cdot}89, 0{\cdot}95$$

for the apples and:

$$0{\cdot}55, 0{\cdot}95, 0{\cdot}80, 0{\cdot}57, 0{\cdot}80, 0{\cdot}79$$

for the pears. It appears then that two years (1948 and 1951) were remarkable for low 'communalities', i.e. for the operation of factors peculiar to themselves. Nor is this surprising, because the first was noteworthy for storms that damaged the fruit on the trees, and the second was abnormally wet during the summer. Hence, the 'communalities' are both consistent and reasonable.

Using these 'communalities' the vectors for the apples become:

$$\zeta_1 = (0{\cdot}48, 0{\cdot}46, 0{\cdot}45, 0{\cdot}25, 0{\cdot}49, 0{\cdot}21)$$
$$\zeta_2 = (0{\cdot}18, 0{\cdot}48, -0{\cdot}38, 0{\cdot}09, -0{\cdot}28, 0{\cdot}72)$$
$$\zeta_3 = (-0{\cdot}05, 0{\cdot}11, -0{\cdot}32, 0{\cdot}86, -0{\cdot}04, -0{\cdot}38)$$

of which the first two account for 87 per cent of the variation and the third for virtually all the rest. For pears they become:

$$\zeta_1 = (0{\cdot}37, 0{\cdot}49, 0{\cdot}39, 0{\cdot}38, 0{\cdot}43, 0{\cdot}38)$$
$$\zeta_2 = (0{\cdot}12, -0{\cdot}38, 0{\cdot}56, -0{\cdot}23, -0{\cdot}45, 0{\cdot}53)$$

accounting for 72 and 22 per cent of the variation respectively. In each instance the first vector plainly represents a general cropping factor. The second vector for the apples might represent a biennial effect with only a small positive effect in the abnormally wet year; for the pears the second factor appears to be a biennial one with a change of phase after the wet year. The factors are therefore at least as reasonable as those found by the component analysis.

Conclusions

It would appear, from these examples, that the method of component analysis applied to problems of fruit tree cropping, growth, and form can give results which are both reasonably consistent and intelligible within the context of that field of investigation.

Apart from being generally more applicable to the study of interrelationships, the correlation matrix, by standardising the variances and covariances, often gives the more readily interpreted vectors. One weakness of the technique, however, is the underlying assumption that all the variation in each variate can be accounted for in terms of the other variates, there being no random error or other sources of variation external to the system. What is required is a technique that allows for this possible variation. The 'communalities' of a factor analysis may themselves be biologically informative but certain *a priori* knowledge is needed to set up a model able to cope with the maximum number of factors likely to arise. It does appear, therefore, that multivariate methods can form a useful extension of the more accepted methods and can lead to a better understanding of the tree *as a whole*. This is essential if the statistician is really to enter into the thought of the biologist, and this, in our experience, he cannot do if he confines himself to univariate methods.

REFERENCES

[1] Holland, D. A. (1959). 'Studies in the measurements of apple trees. IV. The use of branch girths to estimate tree size', *Rep. E. Malling Res. Sta. for 1958*, 77.

[2] Jolly, G. M. and Holland, D. A. (1958). 'Sampling methods for the measurement of extension growth of apple trees', *Rep. E. Malling Res. Sta. for 1957*, 87.

[3] Jowett, G. H. (1958). 'Factor analysis', *Applied Statistics*, **7**, 114.

[4] Kendall, M. G. (1957). *A Course in Multivariate Analysis*. Charles Griffin and Co. Ltd., London.

[5] Pearce, S. C. (1952). 'Studies in the measurement of apple trees. I. The use of trunk girths to estimate tree size', *Rep. E. Malling Res. Sta. for 1951*, 101.

[6] *Idem* (1956). 'Studies in the measurement of apple trees. III. The use of branch girths as a substitute for the trunk girth', *Rep. E. Malling Res. Sta. for 1955*, 76.

[7] *Idem* (1959). 'Some recent applications of multivariate analysis to data from fruit trees', *Rep. E. Malling Res. Sta. for 1958*, 73.

[8] Pearce, S. C. and Davies, G. E. (1954). 'Studies in the measurement of apple trees. II. The height at which trunk girths should be measured'. *Rep. E. Malling Res. Sta. for 1953*, 138.

SOME STATISTICAL ASPECTS OF PREFERENCE AND RELATED TESTS

C. I. BLISS
The Connecticut Agricultural Experiment Station and Yale University, New Haven

In the first part of this paper Dr Bliss outlines the nature and purpose of preference tests and, without claiming to give a complete survey, discusses some suitable experimental designs. He then points out the advantages of analysing separately the results from subjects with consistent preferences and those from subjects with no consistent preferences. Finally, he describes statistical methods for the study of sensory thresholds.

Introduction

In an extended, though not exhaustive, survey of the recent literature, sensory tests were found to range from the checking of individual preferences in questionnaires to ratings by experts, as in the judging of tea or of milk (Fenton[16]). The many procedures that these represent are far more numerous and varied than can be considered here; so the present paper will be limited to comparative sensory experiments in which the subject reports his preferences, and to pertinent related tests. Even with this limitation, the choice of experimental designs, statistical techniques, and citations from the literature represents a personal selection from among the many more that might be mentioned.

The subjects in these experiments consist commonly of such students or staff members as can be induced to participate, more often than not with little or no prior training. When the preferences of the population from which our subjects are drawn are as critical as the materials which they compare, a larger sample that is demonstrably representative of a wide population of potential consumers may be required. Yet a smaller group, if it has not been selected on the basis of its preferences, may also be an unbiased sample of a much larger sector of the consuming public than is suggested by its occupational status. Although some investigators would restrict the term 'preference tests' to the experiments with larger, more representative, samples, the size of the group and its source will not be considered a limitation here.

Many sensory experiments are designed specifically for discriminating differences rather than for estimating preferences. Triangular and duo-trio designs, for example, are explicitly of this type. In each case the subject is presented with three similar items, two of them identical. In the triangle test the odd sample is to be selected; in the duo-trio test one control sample is so designated and the subject is to select the other control sample from the remaining two. Although less efficient per test unit than other designs (Gridgeman[21]), experiments such as

these are well adapted to distinguishing the 'don't care' from the 'can't tell' subjects in a preference test, if the two groups need to be separated.

We shall consider first a few of the experimental designs that are of value in the testing of preferences. Two desirable objectives in a preference experiment are (1) to determine what proportion of subjects has a consistent choice between alternatives, and (2) to measure the direction and degree of preference of the consistent subjects. An experiment may answer both questions, at least approximately, even though it does not separate the subjects who are insensitive from those who are indifferent, as in our second section. Although some designs can be adapted for testing both differences and preferences, our primary concern here will be with preferences and the factors upon which they may depend. One of these factors is the relation between the sensory threshold for a critical component in the population and the concentration of this component in the test materials, which is our last topic.

The Experimental Design of Preference Tests

Experimental preference tests should be restricted to items in the same category, such as different kinds of fruit or different varieties of the same kind of fruit. Since judgements are comparative, each item need not be rated separately in the five or more steps of a hedonic scale, such as: dislike very much, dislike somewhat, neither dislike nor like, like somewhat, and like very much. Individual ratings of this type complicate the comparison of relative preferences unnecessarily by including differences between subjects, both in their average response and in the spread of their answers over the rating scale.

Within these restrictions, what kinds of comparative tests are suitable for preference studies? The choice between two items raises no problem. If we have three or more, one plan is to present them in pairs, so that the subject compares each variety, say, separately with every other variety in a so-called paired comparison (Bradley,[11] Jackson and Fleckenstein[25]). Given four varieties A to D, each subject would judge them in six pairs: A-B, A-C, A-D, B-C, B-D, and C-D. As each pair is presented, he might be asked additionally, as first proposed by Scheffé, 'Is your preference slight, moderate or strong, or in reality non-existent?' (Scheffé,[33] Bliss *et al.*,[10] Carroll[13]). Although differences between subjects in the range covered by their preferences may introduce an element of heterogeneity that is avoided when the degree of preference is not asked, the gain in information will often justify this risk.

An alternative design is to ask the subject to rank the three or more items from the most to the least preferred (Bliss *et al.*,[8, 9] Greenwood and Salerno[20]). To avoid bias, the order of presentation can be randomised separately for each subject or balanced with a Latin square. Since the design should facilitate comparisons within a selected set of stimuli, ranking is most effective when these are qualitatively similar.

If the critical stimuli are qualitatively dissimilar, the subject may have appreciably less difficulty in choosing between the two members of a pair than in ranking three or more in order. Under these circumstances paired comparisons are more likely to reveal inconsistencies in the ranking of stimuli, and for this additional reason would be preferred (Kendall[27]). If ranking a series of more than five or six varieties or 'treatments' would reduce a subject's ability to discriminate, and the testing of all possible pairs would require too many replicates, ranking in balanced incomplete blocks may be the solution (Hopkins,[24] Murphy *et al.*,[30]). With one class of these, the Youden square (Youden[34]), both differences between subjects and the order of presentation are balanced concurrently. A simplified calculation for incomplete block designs has been described by Dykstra;[15] alternatively, ranks may be transformed to mean normal deviates (Fisher and Yates,[19] Table XX) for analysis.

The scope and efficiency of many preference tests can be enlarged by a factorial design of the treatments. In a 2×2 factorial, for example, American and Dutch process cocoa syrups were prepared by both the 'hot' and 'cold' methods and the four combinations were tasted in a paired comparison by each of 30 subjects (Reid and Becker[32]). Only the interaction proved significant, and this led to the recommendation that the 'hot' method be used for American cocoa and the 'cold' method for Dutch cocoa. Other factorial taste tests include a 2×3 palatability test on kale (Greenwood and Salerno[20]) and a $4 \times 2 \times 2$ taste test on jam (Gridgeman[22]). As the number of factors is increased, the experiment may be kept within bounds by a fractional factorial, as shown by Carroll,[13] who studied five formulation variables in the preparation of a pudding, each at two levels. By selecting a particular set of 16 from the 32 possible factor combinations and by using a Scheffé rating for the degree of preference, she measured the effect of each variable by an ingenious graphical technique, with half the number of combinations required for a full factorial.

Ideally, the subjects in a preference test are a random sample from the population whose preferences we are trying to measure. In some experiments this ideal can be approximated, as in the mass consumer evaluation of five ice cream samples at a State Fair reported by Pangborn *et al.*[31] In a k-visit method of consumer testing, Ferris[17] develops a trinomial model for comparing two products (a and b) with real consumer preferences in the proportions π_a, π_b, and π_o, the third fraction representing those unable to discriminate or with no preference. By testing each of many subjects on two or more occasions, he can estimate each proportion and its standard error from the numbers observed for each possible combination of judgements. More often, we are limited to personnel in the testing laboratory, who in time may become practised subjects. Training sessions have so increased the consistency and sensitivity of the ratings in some sensory tests (Bennett *et al.*[4]), that a practice session may be warranted even in large-scale

preference testing. Among the studies on panel selection are comparisons with consumer surveys (Murphy *et al.*[29]), the sequential testing of prospective subjects (Bradley,[11] Armitage[1]), and the relation of screening tests with water solutions to judging ability for foods (Mackay and Jones[28]).

Separation of Consistent and Inconsistent Subjects

Granted that the experimental procedure is satisfactorily controlled, the test materials are uniform, and that motivation is adequate, the subjects in a preference test may represent two groups, those with and those without preferences. If we could estimate the relative proportions in the two groups, this in itself would be a desirable statistic. The subjects without preferences may either be unable to discriminate between the test materials, which seem to them all alike (the 'can't tell' group), or able to discriminate but have no consistent preference for one over another (the 'don't care' group). Some aspects of this second dichotomy will be considered later. Those who do have preferences, of course, need not agree with one another, but our ability to evaluate their relative likes and dislikes may depend upon our success in analysing their responses separately from those of subjects with no consistent preferences.

The relevance of such a separation is well illustrated by a clinical experiment in which three different drug combinations and a *placebo* were compared as headache remedies (Jellinek[26]). Tablets of the four preparations, identical in colour, shape, size, and taste were given to headache-prone patients in successive two-week periods, under concealed code designations, in an order determined by a Latin square. Each patient was to take a tablet every time he developed a headache and to record whether his headache was relieved within half an hour. When the results were analysed, the three active preparations gave relief significantly more often than the *placebo* but did not differ from one another. The frequency distribution of the number of successes reported for the *placebo*, however, proved to be U-shaped. On the hypothesis that patients who never reacted to the *placebo* had physiological headaches and those who were relieved by the *placebo* on one or more occasions had psychological headaches, Jellinek divided his subjects into two series, numbering 79 and 120 respectively. The *placebo* non-reactors discriminated between the three true drugs significantly, but the *placebo* reactors gave all three drugs the same success rate as the *placebo* itself (0·86), from which he concluded that 'discrimination among remedies for pain can be made only by subjects who have a pain on which the analgesic can be tested'.

In applying this principle to the testing of preferences, one approach is that developed by Ferris.[17] Another is to replicate the test with each subject and then test his mean ratings against an error term based upon his own inconsistency. The direction of preference or its additivity would *not* be a criterion, only the requirement that a subject designated

as 'having preferences' must show some stability in at least one choice in replicated tests. For those without preferences, all variances between 'treatments' should be of the same magnitude as the error; but for those classed as having a preference, one or more comparisons should exceed the error significantly or approach significance. Since the relative size of the two groups would itself be one outcome of the experiment, we could minimise or omit any preliminary screening of subjects. Individuals with preferences may differ to such an extent that no direct comparisons between treatments are significant, but in this case a significant interaction of treatment by subject would testify to a difference in opinion as contrasted with no opinion.

As an example, this procedure may be applied to a paired comparison study of the relative palatability of Cortland apples from trees which had been treated with four different spray combinations in a single experimental orchard (Bliss *et al.*[10]). Each spray mixture contained one of the two fungicides, thiram (Th) or sulphur, and one of the two insecticides, lead arsenate (L) or parathion, in a 2×2 factorial design. Apples from each treatment were chosen at random from the fall harvest, washed in detergent suds, rinsed, cored, and quartered, made into sauce and quick-frozen. Twenty-five subjects, students and faculty at the University of Connecticut, participated in six replicate taste sessions spread over a period of 26 days. The six possible pairs of the four treatments, all with concealed identities, were presented to each taster in an order determined by a 6×6 Latin square, each subject recording both his preference within each pair and its degree as large, moderate or slight, or the non-existence of preference.

The degrees of preference have been transformed to mean normal deviates or rankits for $N = 7$, and analysed in terms of Scheffé's extension of the Thurstone-Mosteller model. This model postulates a subjective continuum within each subject on which the sensations developed by the stimuli are arranged on a linear scale, the sensations for each stimulus varying normally about a mid-point. The six rankits in each of the six replicated tests for each subject, one for each pair of samples, vary about zero with 36 degrees of freedom; six of these represent differences between the means of the paired stimuli and 30 the remaining variability. From the six mean differences we can isolate three factorial comparisons, each with one degree of freedom, representing the preference between the fungicides thiram and sulphur, between the insecticides lead arsenate and parathion, and their interaction. The sum of squares for the remaining three degrees of freedom measures non-additivity on the subjective continuum, or 'deviations from subtractivity' as they are termed by Scheffé.[33]

Each factorial effect for each subject, disregarding the direction of the preference, was compared against his residual variability. In 11 of the 25 subjects, no effect was significant at $P \leqq 0{\cdot}15$, with $P < 0{\cdot}20$ for only five of the 33 factorial comparisons and no two of these in the same individual. These subjects apparently were either indifferent or

insensitive to any specific flavours which might be associated with the four toxicants. Each of the other 14 reacted with some measure of consistency, even though so slight that one or more might well shift to the no-preference group if the experiment were repeated.

TABLE I[10]

Analysis of variance of a 2 × 2 factorial experiment on off-flavour in apples sprayed with thiram (Th) or sulphur and with lead arsenate (L) or parathion; from paired comparisons with degrees of preference transformed to rankits

Row No.	Comparison of:	Inconsistent Subjects DF	MS	F	Consistent Subjects DF	MS	F
1	Fungicides (Th)	1	0·545	2·22	1	8·417	7·42*
2	Insecticides (L)	1	0·239	0·97	1	3·740	4·91*
3	Interaction Th × L	1	0·004	0·02	1	0·681	3·07
4	Non-additivity	3	0·168	0·68	3	0·020	0·09
5	Tasters × Th	10	0·193	0·78	13	1·134	5·10**
6	Tasters × L	10	0·167	0·68	13	0·761	3·43**
7	Tasters × Th × L	10	0·284	1·16	13	0·132	0·59
8	Tasters × non-add.	30	0·345	1·40	39	0·164	0·74
9	Within tasters	330	0·246		420	0·222	

*$P < 0{\cdot}05$, **$P < 0{\cdot}001$

A separate analysis of variance has been computed for the 11 'inconsistent' and for the 14 'consistent' subjects, and their respective mean squares are given in Table I. In agreement with our hypothesis, no comparison in the first group is significant, although the contrast between thiram and sulphur is twice its error variance ($F = 2{\cdot}22$). By contrast, the other 14 subjects differ significantly in their preferences for the two direct factorial comparisons (rows 5 and 6), but their disagreement is not sufficient to preclude a significant 'vote' for thiram in preference to sulphur ($P < 0{\cdot}025$) and for lead arsenate in preference to parathion ($P < 0{\cdot}05$) in rows 1 and 2, where each direct effect has been compared against its interaction with tasters. The interaction in row 3, measuring the dependence of the preference for the insecticide upon which fungicide was present (or vice versa), approaches significance. The assumption of an additive scale or linear subjective continuum is justified by variance ratios (F) in rows 4 and 8 that are less than 1, i.e. not significant. In comparison with the original composite analysis, the present subdivision of the 25 subjects into two groups has sharpened our test of the disagreement between those with a consistent preference, and has shifted the experimental error for each direct effect to its interaction with subjects.

Lack of Preference and the Sensory Threshold

After the subjects with preferences have been separated from those without, the latter group is still a composite of individuals who can discriminate but have no preference and of those to whom all treatments seem alike. These could, perhaps, be separated by a suitably designed experiment in which the subjects were first asked to detect a difference and then to express a preference, although 'difference' tests

for sensory discrimination may follow quite different patterns from those for measuring preferences, as has been noted. For a better understanding of the relations underlying both types of test, the distribution of the threshold of perception for a critical component may be studied separately.

Individuals who have no consistent preference because they cannot distinguish between the several alternatives may be limited in two different ways. All concentrations of the critical component may fall below the subject's threshold of perception, so that his choices are essentially random. Alternatively, the difference between two concentrations of a critical component, where both are above his threshold, may not exceed the just perceptible difference between them, so again his choice is random. Although both factors may be involved, we shall consider here only the threshold of perception and its distribution. This threshold may vary continuously or discontinuously in a population of subjects.

(*a*) *Continuous Distributions of Sensory Thresholds.* Let us suppose that we are determining preferences for samples of a soft drink, which differ primarily in their content of sugar or of acid. We will further assume that the differences between samples exceed the just perceptible sensory difference, if the subject can taste the sugar or acid at all. A subject with a threshold below that of the sample with the smallest concentration will be able to distinguish one drink from another and record his preference if he has one. A subject with a sensory threshold above the concentration in all or most sample drinks will be unable to separate those falling below his threshold, and will rank them at random or by some secondary factor. Hence, a knowledge of the distribution of the sensory thresholds for primary taste sensations, such as sweet and sour, in a group of tasters would aid in interpreting any preferences in which they may be critical.

Two experiments in this area have concerned water solutions of sucrose (Baker *et al.*[2]) and of tartaric acid (Baker *et al.*[3]) differing by steps of two in a geometric progression. Each subject was asked to identify the test solution when it was paired against water, and he was presented with each concentration in his critical range 15 or more times over a number of sessions. Concentrations below his sensory threshold have an expectation of 50% correct answers, which increases to 100% as the concentrations reach and exceed the subject's threshold. The statistical problem is two-fold: (1) how can we compute the median threshold concentration for each subject; and (2) how are these thresholds distributed in the population from which the subjects may be considered a random sample?

The smallest concentration which can be detected is a measure of the threshold level at the time of the test, although in replicate trials with the same concentration the subject may react on some occasions and not on others. In repeated trials, with concentrations ranging from one which the subject cannot distinguish from the control to

one which he always identifies correctly, the percentage of positive reactions plots against the log-concentration as a symmetrical sigmoid curve, as has been shown for sucrose by Gridgeman.[23] If the variation in the threshold results from a number of factors acting concurrently, some raising and others lowering the level, a suitable model would interpret the sigmoid relation as an integrated normal curve.

Taste tests have the complication that we start with an expectation of 50 instead of 0 per cent of correct answers at concentrations where the subject is unable to discriminate. The problem has long been known and solved for insecticide experiments in which natural mortality and the toxicant act independently of each other. With an expectation equivalent to a known natural mortality of 50%, each initial response between 50 and 100% of correct identifications in a taste test can be converted to a

$$\text{Net \% response} = \frac{\text{observed\%} - 50}{1 - 0{\cdot}50} = 2(\text{observed\%} - 50)$$

At the smaller concentrations in a graded series, percentages of more than 50% that are not followed by larger values may exceed 50% by chance and can often be omitted as irrelevant.

For analysis, the sigmoid curve is converted to a straight line by transforming each net response in the intermediate zone between 0 and 100 per cent to its empirical probit by a suitable table (Fisher and Yates[19]), the empirical probits for responses of 0 and 100 per cent being minus and plus infinity respectively. When there are few concentrations in the intermediate zone, as in the present series with sucrose and with tartaric acid, we may improve our trial estimates by a simple dodge (Berkson[5]). The first net 0% below the intermediate zone and the first 100% above the intermediate zone are replaced provisionally by the percentages $100/N$ and $100(N-1)/N$ respectively, where N is the total number of pairs sampled by a single subject at a given concentration. These percentages are then transformed to empirical probits.

Given the coded log-concentrations (x) and the corresponding empirical probits (y), we may compute a trial straight line for each individual by unweighted least squares. When the slopes (b) of two or more lines agree substantially, they can be combined into improved, more stable estimates of the slopes b_c; in Baker's experiments, for example, the subjects tested with sucrose fell into two homogeneous groups and those tested with tartaric acid into three groups. Each individual's threshold in coded log-units may be estimated provisionally from the unweighted means, $\bar{x}$ and $\bar{y}$, and from b_c as

$$X_5 = \bar{x} + (5 - \bar{y})/b_c$$

where X_5 is the coded mean of a log-normal sample with an estimated standard deviation of $s = 1/b_c$. The calculation of the provisional

estimate is illustrated in Table II for tests with sucrose on subjects 'C' and 'N'.

For a definitive result, these initial estimates are improved by maximum likelihood, replacing the empirical probits by their estimated working probits y, and weighting each y with a weighting coefficient for 50% natural mortality (Finney[18]). Agreement with our model can then be checked by χ^2. A composite χ^2 over Baker's tests with sucrose and with tartaric acid showed adequate agreement with the underlying hypothesis ($\chi^2 = 82{\cdot}77$, $n = 75$).

TABLE II[2]

Provisional estimate of EC50 for tasters 'C' and 'N' from paired difference tests against water of five concentrations of sucrose increasing by multiples of two from 0·034% at $x=1$; x is the coded log-concentration

Taster	Solution x	No. Correct / Total	Net %	Probit y	Calculation in Coded Log-concentrations:
'C	1	8/19	0	3·4	$\bar{x}=2{\cdot}5$, $\bar{y}=4{\cdot}575$
	2	10/18	11	3·8	$[x^2]=5$, $[xy]=5{\cdot}05$
	3	12/18	33	4·6	Slope for 'C', $b=1{\cdot}01$
	4	15/15	100	6·5	Pooled slope $b_c=1{\cdot}120$
	5	15/15	100	Omit	EC50=2·879 (using b_c)
'N'	1	10/19	5	3·4	$\bar{x}=3$, $\bar{y}=5{\cdot}32$
	2	13/17	53	5·1	$[x^2]=10$, $[xy]=7{\cdot}2$
	3	16/19	68	5·5	Slope for 'N', $b=0{\cdot}72$
	4	14/15	87	6·1	Pooled slope $b_c=0{\cdot}725$
	5	15/15	100	6·5	EC50=2·559 (using b_c)

Weighted estimates of EC50 $= X_5 = 3{\cdot}074$ for 'C' and 2·377 for 'N'.

From the weighted means $\bar{y}$ and $\bar{x}$, the coded log-threshold-concentration X_5 for sugar or acid was recomputed for each subject with the relevant combined slope b_c. The agreement of these log-thresholds X_5 with the normal curve could be examined by a simple graphic test. Each series has been arranged in increasing order and plotted in Fig. 1, where the ordinate is the corresponding rankit, or expected normal deviate for samples of $N = 15$ and 24 (Fisher and Yates,[19] Table XX). The trend of each series of plotted points is here substantially linear, but with the standard deviation for tartaric acid 2·6 times that for sucrose. Since this graphical result is consistent with the distributional pattern that recurs continually in a great variety of experiments with thresholds, we may consider the distribution of these two taste thresholds as similarly log-normal. If low concentrations of sucrose or tartaric acid were critical in samples of a soft drink offered to a panel, they could influence the choice of only those individuals with a small enough threshold to taste these components.

(*b*) *Discontinuous Distributions of the Sensory Threshold.* The variation of sensory thresholds between individuals may not be continuous and substantially normal (Gaussian), as in the above tests with sucrose and tartaric acid. Instead a clear-cut discontinuity may divide the population into two categories of tasters and non-tasters. Perhaps the best known case is that of solutions of phenyl-thio-carbamide (PTC), which to some people are exceedingly bitter and to others tasteless.

Tasters can be separated from non-tasters by a solution of about 1/5 molar. In a study of some 3700 individuals, the proportion of tasters

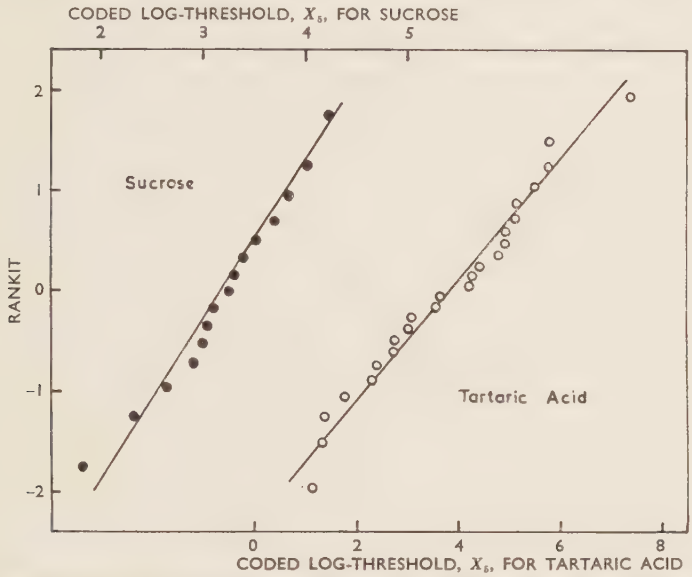


FIG. 1. Graphical test for agreement with a normal distribution of the coded log-threshold-concentrations X_5 in 15 subjects for sucrose and in 24 subjects for tartaric acid. Successive concentrations increase by doubling from $1 \equiv 0{\cdot}034\%$ of sucrose and $1 \equiv 586 \times 10^{-8}$ moles per litre of tartaric acid; at rankit 0, the straight lines pass through means of $\bar{X}_5 = 3{\cdot}210$ for sucrose and $\bar{X}_5 = 3{\cdot}834$ for tartaric acid with slopes of $b = 1/s$, where the standard deviation $s = 0{\cdot}635$ and $1{\cdot}651$ respectively.

in the population was about 71% (Cotterman and Snyder[14]). Geneticists have traced the dichotomy to a single pair of alleles. Individuals homozygous for the recessive gene find PTC tasteless and individuals with its dominant allele find it extremely bitter. Although in saturated solutions (4/5 molar) all individuals can detect some bitterness (Blakeslee[6]), the frequency distribution of the taste threshold is sharply bimodal, with a fair spread between tasters and little or no spread between the non-tasters.

Blakeslee and his associates[7, 12] have extended their studies to olfactory as well as taste reactions. Individuals varied widely not only in their thresholds but also in their opinions whether a given odour was pleasant, indifferent, or unpleasant. If the preference for one of two varieties of a given food, for example, were to depend upon a genetic bimodality in the taste or olfactory threshold, the situation would parallel that separating the *placebo* reactors from the *placebo* non-reactors in Jellinek's experiment with headache remedies. The indifferent subjects who are also insensitive may represent 'taste blind' individuals, as in the PTC test, or the high-threshold end of a Gaussian distribution. In the latter case, we might expect the proportion of non-tasters to be relatively unstable and more dependent upon the exact concentration of the critical component in the test materials than if the distribution of sensory thresholds were bimodal.

REFERENCES

[1] Armitage, P. (1957). 'Restricted sequential procedures', *Biometrika*, **44**, 9.

[2] Baker, G. A., Amerine, M. A. and Roessler, E. B. (1954). 'Errors of the second kind in organoleptic difference testing', *Food Research*, **19**, 206.

[3] Baker, G. A., Mrak, V. and Amerine, M. A. (1958). 'Errors of the second kind in an acid threshold test', *Food Research*, **23**, 150.

[4] Bennett, G., Spahr, B. M. and Dodds, M. L. (1956). 'The value of training a sensory test panel', *Food Technology*, **10**, 205.

[5] Berkson, J. (1953). 'A statistically precise and relatively simple method of estimating the bio-assay with quantal response based on the logistic function', *J. Amer. Statist. Assoc.*, **48**, 565.

[6] Blakeslee, A. F. (1932). 'Genetics of sensory thresholds: taste for phenyl-thiocarbamide', *Proc. Nat. Acad. Sci.*, **18**, 120.

[7] Blakeslee, A. F. and Salmon, T. N. (1935). 'Genetics of sensory thresholds: individual taste reactions for different substances', *Proc. Nat. Acad. Sci.*, **21**, 78.

[8] Bliss, C. I., Anderson, E. O. and Marland, R. E. (1943). 'A technique for testing consumer preferences, with special reference to the constituents of ice cream', *Storrs Agric. Expt. Sta., Bull.* 251, 1.

[9] Bliss, C. I., Greenwood, M. L. and McKenrick, M. H. (1953). 'A comparison of scoring methods for taste tests with mealiness of potatoes', *Food Technology*, **7**, 491.

[10] Bliss, C. I., Greenwood, M. L. and White, E. S. (1956). 'A rankit analysis of paired comparisons for measuring the effect of sprays on flavor', *Biometrics*, **12**, 381.

[11] Bradley, R. A. (1953). 'Some statistical methods in taste testing and quality evaluation', *Biometrics*, **9**, 22.

[12] Campbell, W. I. P. and Blakeslee, A. F. (1948). 'Would a rose smell so sweet by any other name', *Horticulture*, **24**, 333.

[13] Carroll, M. B. (1958). 'Consumer product testing statistics', *Flavor Research and Food Acceptance*, pp. 162-174. Reinhold Publ. Corp., New York.

[14] Cotterman, C. W. and Snyder, L. H. (1939). 'Tests of simple Mendelian inheritance in randomly collected data of one and two generations', *J. Amer. Statist. Assoc.*, **34**, 511.

[15] Dykstra, O. (1956). 'A note on the rank analysis of incomplete block designs—applications beyond the scope of existing tables', *Biometrics*, **12**, 301.

[16] Fenton, F. E. (1957). 'Judging and scoring milk', *Farmer's Bull.* 2111, 1.

[17] Ferris, G. E. (1958). 'The *k*-visit method of consumer testing', *Biometrics*, **14**, 39.

[18] Finney, D. J. (1952). *Probit Analysis*, 2nd Ed. Cambridge University Press.

[19] Fisher, R. A. and Yates, F. (1957). *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edinburgh and London.

[20] Greenwood, M. L. and Salerno, R. (1949). 'Palatability of kale in relation to cooking procedure and variety', *Food Research*, **14**, 314.

[21] Gridgeman, N. T. (1955). 'Taste comparisons: two samples or three?', *Food Technology*, **9**, 148.

[22] Gridgeman, N. T. (1956). 'A tasting experiment', *Applied Statistics*, **5**, 106.

[23] Gridgeman, N. T. (1958). 'Application of quantal response theory to the cross-comparison of taste-stimuli intensities', *Biometrics*, **14**, 548.

[24] Hopkins, J. W. (1954). 'Incomplete block rank analysis: some taste test results', *Biometrics*, **10**, 391.

[25] Jackson, J. E. and Fleckenstein, M. (1957). 'An evaluation of some statistical techniques used in the analysis of paired comparison data', *Biometrics*, **13**, 51.

[26] Jellinek, E. M. (1946). 'Clinical tests on comparative effectiveness of analgesic drugs', *Biometrics Bull.*, **2**, 87.

[27] KENDALL, M. G. (1955). 'Further contributions to the theory of paired comparisons', *Biometrics*, **11**, 43.

[28] MACKAY, A. O. and JONES, P. (1954). 'Discernment of primary tastes in water solution compared with judging ability for foods', *Food Technology*, **8**, 527.

[29] MURPHY, E. F., CLARK, B. S. and BERGLUND, R. M. (1958). 'A consumer survey versus panel testing for acceptance evaluation of Maine sardines', *Food Technology*, **12**, 222.

[30] MURPHY, E. F., COVELL, M. R. and DINSMORE, J. S., JR. (1957). 'An examination of three methods for testing palatability as illustrated by strawberry flavor differences', *Food Research*, **22**, 423.

[31] PANGBORN, R. M., SIMONE, M. and NICKERSON, T. A. (1957). 'The influence of sugar in ice cream. I. Consumer preferences for vanilla ice cream', *Food Technology*, **9**, 679.

[32] REID, A. W. and BECKER, C. H. (1956). 'A study of cocoa syrups for taste preference', *J. Amer. Pharm. Assoc., Sci. Ed.*, **45**, 160.

[33] SCHEFFÉ, H. (1952). 'An analysis of variance for paired comparisons', *J. Amer. Statist. Assoc.*, **47**, 381.

[34] YOUDEN, W. J. (1940). 'Experimental designs to increase accuracy of greenhouse studies', *Contrib. Boyce Thompson Inst.*, **11**, 219.

A STUDY OF SOME POTENTIAL BIASES IN THE OPERATION OF A CONSUMER PANEL*

A. S. C. EHRENBERG†
The Attwood Group

Mr Ehrenberg here describes some controlled experiments with consumer panels in which the effects of length of panel membership and of the number of product-fields covered were studied.

Introduction

The possible occurrence of biases in operations of the panel type is a matter of frequent concern to those engaged in market and social research. The biases occurring in the selection of panel-members is, for example, one of the current research topics of the Survey Research Unit at the London School of Economics. Wadsworth,[1] in describing in this journal his experiences as a user of a consumer panel, discussed among other matters the over-statement of consumer purchases for the products concerned; this is a problem of particular difficulty, as such over-statement has not been found to be the general rule with consumer panels. Again, Prais,[2] in a recent paper read before the Royal Statistical Society, presents evidence of various biases in cost-of-living or household budget enquiries where the informants were asked to record, for periods of a month or two, all items of expenditure, or at least all expenditure on certain very broad product-groups such as food or clothing.

In the present paper, experimental evidence is given on two possible sources of biases in the operation of consumer panels such as those run by Attwood's on a more or less continuous basis, where purchasing records are required for a restricted range of specific commodities.

With a continuous consumer panel, the length of any person's panel-membership necessarily increases as time goes on, and it is a major problem whether the length of panel-membership becomes a source of bias in the recording of purchases. A second point of concern is whether increasing the number of commodities for which purchases have to be recorded will affect the quality of the purchasing records obtained. Before the experimental data are presented it is, however, necessary to sketch in the relevant aspects of the consumer panel technique that is operated by the Attwood Group.

* Based on a paper read at the Cambridge Conference of the General Applications Section of the Royal Statistical Society, September 1959.

† Now at Research Services Limited.

The Consumer Panel Method

An Attwood Consumer Panel is based on a stratified multi-stage random sample of housewives. If left to itself, such a sample would become unrepresentative in the course of time, as the age of panel-members increases and other changes in the composition of the selected households occur. Furthermore, some selected housewives fail to co-operate in the first place, and others will remove themselves from the panel after a time. All this tends to happen more for some types of housewives than for others (e.g. younger housewives tend to co-operate more readily, and a striking geographical differential was recently found in Italy, where the refusal rate in the less sophisticated Southern parts was very substantially lower than that in the larger towns of Northern and Central Italy).

For these various reasons, and also because a random sample can in any case be improved by stratification, a detailed stratification or control by demographical and geographical factors is used in the sampling of the housewives themselves. In effect, the procedure is to establish a target number for each stratum and then to sample the required number of housewives in that stratum at random. Thus the Attwood Consumer Panel in Great Britain is controlled in 1008 strata, based on size of household, age of housewife, social class, presence of children, and geographical regions. (The targets in over a third of these strata are zero to the nearest whole number when one is dealing with a panel of the order of 2000 housewives.) The composition of the panel is further controlled according to the target numbers for each sampling district which are prescribed by the multi-stage sample design and by factors such as television ownership and purchasing habits at co-operative outlets. In practice, slight temporary deviations from these various targets are at times unavoidable, but by special panel control procedures their effect is minimised. In this way, it is possible in a very large measure to keep the composition of a consumer panel representative and continuously balanced.

The weekly 'diary' in which the member of an Attwood Consumer Panel has to record her purchases covers a selection of non-durable consumer goods, mainly in the branded goods market, typified by the following list:

> Butter, Cocoa and Drinking Chocolate, Coffee, Cooking Fat, Cream, Detergents, Dentifrice, Face Tissues, Floor or Furniture Polish, Flour, Food Drinks, Household Soap, Ice Cream, Jam, Margarine, Milk, Packaged Soup, Porridge, Potatoes, Processed or Pre-cooked Meat, Ready-to-Eat Cereals, Salad Cream and Mayonnaise, Salad Oil, Scourers, Shampoos, Shoe Polish, Soft Drinks, Tea, Toilet Soap, and Toilet Tissues.

Product headings, and also headings for other items of information which are required (such as brand and manufacturers' names, how

much was bought, how much was paid, and where and when the purchases took place) are set out in the diary.

The Effect of the Length of Panel Membership

In the course of running continuous consumer panels for periods of up to ten years, Attwoods accumulated many pieces of evidence which indicated that increasing length of panel membership, even when extending over quite a number of years, did not affect the level of reported purchases. For example, it had been found both in this country and in Germany that the purchasing rates of newer panel-members did not differ systematically from those of older members; and that post-mortem analyses of past purchasing records, according to whether or not the panel-member subsequently left the panel, uncovered no general trends. Similarly, long-term comparisons of panel data with such sales or production statistics as were available had shown up no systematic pattern of progressive divergencies. All this evidence was fairly conclusive, but it was not, on technical grounds, entirely water-tight. Recently, however, evidence on the effect of length of panel membership has become available from fully controlled experimental situations in both Holland and Great Britain.

Holland. The Attwood Consumer Panel of 750 housewives in Holland had been running for over three years when it was decided to increase the sample to 1000 housewives in the winter of 1957-8. At the same time, the number of sample districts was increased by about 80%. Some 400 new housewives had to be recruited to effect this double change. This operation provided an opportunity for comparing the purchases recorded by the 'old' panel of 750 housewives, two-thirds of whom had been on the panel for over three years and only about 7% for less than a year, with the purchases of a 'new' panel of 1000 housewives, 40% of whom had been very recently signed up. Both the 'old' and 'new' panels were in other respects directly comparable, being random samples from the Dutch private household population, controlled and balanced as described on page 21. It was not possible to compare the newly-recruited housewives on their own directly with the 'old' ones, because of essential differences in the composition and location of the two samples. In comparing the two panels, about 600 housewives in the two panels were, therefore, identical, but this does not affect the validity of the comparison of 'old' with 'new', although it will tend to bring the results closer together and has to be allowed for in tests of significance. (A difference of little over 10% in the purchasing rates of the old panel and of the new panel-members on their own would have shown up as statistically significant, so that the comparison is a reasonable sensitive one, especially when searching for consistent trends for all products taken together.) The comparative data which are used refer to a four-weekly period in January 1958.

Great Britain. During the winter of 1957-8 the Attwood Consumer Panel of 2000 housewives in Great Britain, which had been running

for almost ten years, was doubled in size by adding to it a second new panel of 2000 housewives. This provided an opportunity, similar to that in Holland, of comparing purchases recorded by an 'old' panel of 2000 housewives, about half of whom had been panel-members for over five years, with a completely new panel of 2000 housewives, both panels being random samples from the British private household population, balanced and controlled as previously described. The comparative data used mainly refer to a twelve-weekly period in the summer of 1958, although for extraneous practical reasons four- or eight-weekly periods are involved in some cases, and in a number of instances 'new' panel data are as yet only available from balanced samples of either 1660 or 1330 households instead of from the full new sample of 2000 households.

Two main comparisons between the 'old' and 'new' panels have been made both in Holland and Great Britain, namely for total purchases and for brand shares, as follows:

First, the total purchases recorded for each commodity analysed have been compared, as shown in Fig. 1. The purchases of any one commodity were initially expressed in terms of weight, or packages, or standard units, or expenditure, whichever was the type of unit that was normally used in analysing that product. However, to bring the results for the different commodities onto a meaningful common scale, they were then indexed by using the typical number of diary entries per week for each commodity as a base, the number of diary entries being a measure of the frequency of purchase. Thus, if the 'new' panel spent 6% more on a given commodity than the 'old' panel and the typical number of diary entries is 1·50, the numbers plotted are 1·50 ('old' panel) and 1·59. On each axis the values are plotted on a square-root scale.

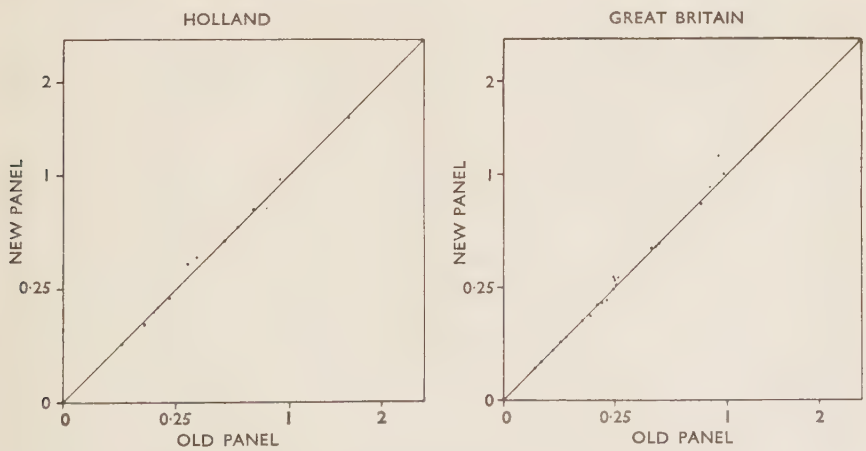


Fig. 1. Purchasing rates of various product fields indexed on the average number of diary entries per week and compared for old and new panels.

Secondly, the market shares held by the various brands of each

commodity have been compared, as shown in Fig. 2. Something like six to eight leading brands were separately itemised for each product, the remaining smaller brands being grouped together as is customary in consumer panel analyses. The brand shares are based either on quantities bought or on expenditure, as the case may be, but are of course directly comparable.

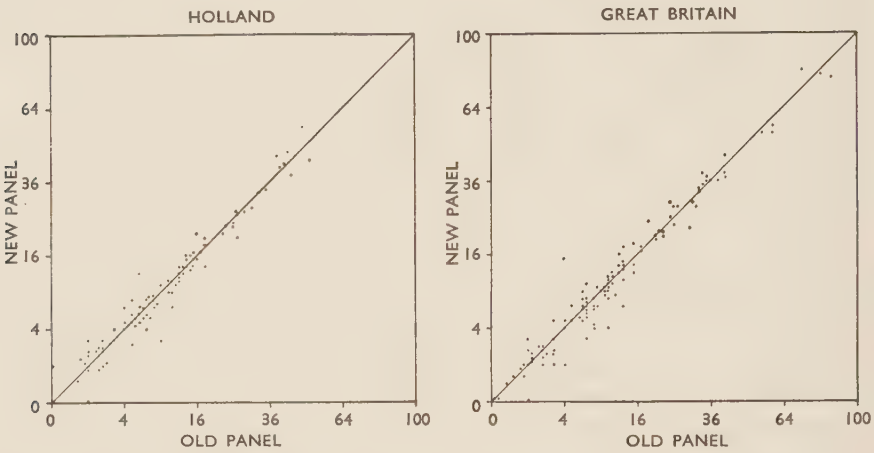


FIG. 2. Brand shares in various product fields compared for old and new panels.

In addition, other forms of comparisons between the 'new' and 'old' panel data have been carried out, but they do not add anything further to the results shown in Figs. 1 and 2. (For example, the percentage of households buying each commodity at all have been compared, and so has the penetration of each brand. Area and other breakdowns have been considered, brand shares have been examined separately for each individual product, and a different measure of frequency of purchase, namely, the average frequency of buying per buyer, has also been used.)

From the results presented in Figs. 1 and 2, we see:

(i) That there is no general systematic bias in the 'old' panel results as against the 'new' one.

(ii) That such differences as exist between the 'old' and 'new' results show no trend either with increasing frequency of purchase of the product or with increasing brand share.

(iii) That the discrepancies between the 'old' and 'new' purchasing rates or brand shares are generally small. In Fig. 1 most differences are less than 5%, with some at about 10% and only one or two still larger ones; few of the brand share differences in Fig. 2 are more than 1 or 2 percentage points. (Individual differences seldom reach the 5% level of significance—the background of sampling error calculations for consumer panel data is given elsewhere.[3])

(iv) That the two sets of experimental results, from Holland and Britain, are in agreement.

Special reasons could be found for some of the larger discrepancies.

Thus, two of the largest discrepancies in Fig. 1 are for products with exceptionally high rates of purchasing per buyer where the sampling errors are relatively high, and in one case the design of the relevant diary entry was known to be still unsatisfactory. Many of the larger discrepancies in Fig. 2 relate to brands of less heavily purchased commodities, with high sampling errors, whilst the essential lack of independence of the brand shares resulted in a decided tendency for a large brand-share discrepancy to be mirrored by a corresponding large discrepancy for another brand in the same product-field. However, despite all these reasons, one or two of the larger discrepancies are still inexplicable, as sometimes happens in market research. But the main point to be made here is that it has not been possible to discover any general pattern in the behaviour or background of these larger discrepancies.

These findings confirm the impressions mentioned earlier that for consumer panels of the kind described, the length of panel membership, even when stretching over a number of years, does not systematically affect the reported results. The reason is presumably that much the easiest and also the most satisfying way of filling in purchasing diaries week after week is to try and do so accurately (assuming that the checks and supervision of the informants which are appropriate to such consumer panel operations are made), and also that any 'conditioning' of the housewives caused by having to record their purchases is likely to wear off or stabilise after a few weeks.

These findings do, however, differ from the conclusions drawn by Prais[2] regarding the existence of 'length of enquiry' and 'recording fatigue' biases. The reason for this may be a technical difference in methods. Thus, in the Attwood panels, the purchasing records obtained during an initial 'running-in period' of six to eight weeks for new panel-members are disregarded for market analysis purposes. (Some over-reporting can then take place, owing to the recording of existing stocks in the home as well as of purchases, or owing to exceptional purchases made either for prestige reasons or because stock-deficiencies have become more apparent than usual, or owing to false reporting for prestige reasons, and so on. In addition, various other reporting mistakes are more likely to occur whilst the panel-member is new.) On the other hand, the diary or account-book data referred to by Prais related only to purchases recorded over periods of two months or less, i.e. the duration of the disregarded running-in period for Attwood Consumer Panel members. The suggestion lies to hand that Prais's supposed 'fatigue' and 'length of enquiry' biases may not have the positive meaning implied, but a negative one, namely, that of a disappearance of initial over-reporting. (See note at end of paper.)

The Number of Commodities Recorded

In consumer panels of the kind described, the effect of increasing the size of the diary and the number of commodities listed in it has

had to be watched carefully. Here also no deleterious effects had in general been observed in the past by Attwoods, despite very marked differences in the size of the diary used at different times and for different panels. However, the change in the sample design of the Dutch consumer panel which has already been mentioned created an opportunity for conducting, at reasonable cost, a fully-controlled experiment on the effect of increasing the number of commodities to be recorded in the diary.

In this experiment a randomised split-run design was used, one half of a sample of 332 panel-members continuing with the standard diary which listed about 30 product-fields, the other half being changed over to a larger diary, listing something over half as many additional commodities on about 50% more page-space and yielding a little more than half as many extra diary entries. The basic procedure was to match the total sample in pairs, according to sampling district and total number of diary entries in an earlier base period, and then to select one panel-member randomly from each pair for the larger diary. (These 'experimental' housewives were discarded after the experiment.) The recorded purchases of the commodities common to the standard and the larger diary were then compared, to see whether these records were affected by having, in the one case, to describe purchases of another set of commodities. The test took place during a period of twelve weeks in the spring of 1958.

The average number of diary entries for each commodity during the last four-weekly period of the experiment recorded by the experimental group is compared in Fig. 3 with that of the control group, the latter

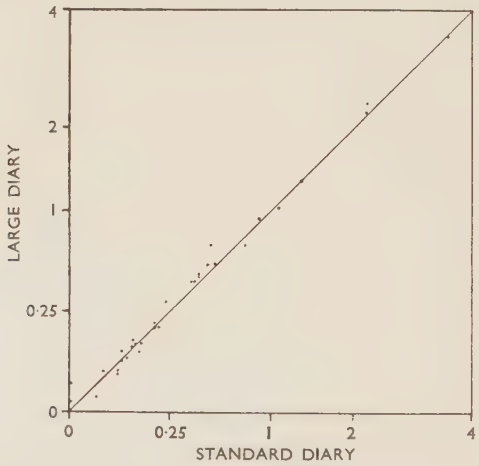


Fig. 3. The average number of diary entries per week for different product fields compared for the large and the standard diaries (Holland).

readings having been adjusted for the differences which occurred between the two groups in the four-weekly period before the larger diary was introduced.

From Fig. 3 we see that there is no consistent difference between the

experimental and control samples, nor any trend with the frequency of purchase. Most of the differences are in fact quite small. In testing the individual differences for significance, only one or two reached the 5% level, which is what one would expect with the number of differences involved.

More detailed investigations of the results have also been made, for example the trends in the number of diary entries during the twelve weeks of the test and the actual quantities bought (rather than the sheer number of entries) have both been scrutinised. However, these investigations also have shown no systematic effects due to the larger diary. The absence of a trend during the twelve weeks is particularly interesting because it indicates, although not perhaps conclusively, that any increased interest stimulated in the experimental group by the introduction of the larger diary—which one might possibly expect to lead to more careful recordings for a few weeks—did not cause an increase in the number of diary entries recorded.

The general conclusion is therefore that an increase in the number of products to be recorded from just over 30 to about 50 does not affect the accuracy of recording. This result concerns housewives who had already been using the small diary for many months or even for two or three years. The negative conclusion here, therefore, does not necessarily contradict Prais's conclusion, namely, that the recording of 'Y' can affect the records of 'X', since the Israeli enquiry on which his conclusion was based lasted for only one month in all.

Note Added in Proof

In the discussion of this paper at the Cambridge Conference, the possible existence of initial over-reporting, as mentioned on page 25, was critically queried. It should, therefore, perhaps be stressed here that there is no really watertight evidence that initial over-reporting on the Attwood type of consumer panel does occur. However, such findings as are available on the initial performance of panel members recruited by the Attwood procedure have not allowed the rather costly six to eight weeks 'running-in' period to be discontinued.

REFERENCES

[1] Wadsworth, R. D. (1952). 'The experience of a user of a consumer panel', *Applied Statistics*, **1**, 169.

[2] Prais, S. J. (1958). 'Some problems in the measurement of price-changes with special reference to the cost of living', *J. R. Statist. Soc.*, A, **121**, 312.

[3] Ehrenberg, A. S. C. (1959). 'The pattern of consumer purchases', *Applied Statistics*, **8**, 26.

AN APPLICATION OF LINEAR PROGRAMMING TO AGRICULTURAL ECONOMICS

SHIRLEY E. VINCENT
Bedford College, Regent's Park, London, N.W.1

In this article it is first shown how a problem concerning the organisation of a dairy enterprise was set out in linear programming terms. Since the requisite analysis would have been impracticable by the Simplex method without an electronic computer, an alternative method suitable for a desk machine is then described.

Introduction

This paper contains an account of work carried out to solve a problem in agricultural economics by the use of linear programming techniques. It was assumed that one section of a farm was to be devoted entirely to a dairy enterprise, and the problem was to find how to organise this section so that its annual net profit would be maximised. As the dairy enterprise was considered to be only a part of the total farm plan, no restrictions were placed on labour and capital available to the enterprise.

Assumptions and Restrictions

It was not desired that there should be any particular acreage available for the enterprise; the interest lay in determining the most profitable layout and intensity of stocking of the farm. For convenience, the calculations were carried out on the basis of 100 acres of available land; farms of other sizes would be planned in proportion.

The herd size was kept constant throughout the year. A yield of 800 gallons per cow per year was assumed, with a pattern of production resulting from two-thirds of the herd calving in November and one-third in April. The average yield per cow per month from this system is given in Table I. This yield and pattern are typical of East Anglian dairy farms.

Each cow requires a minimum of 7 lb starch and 0·75 lb protein per day for maintenance, and an extra 2·5 lb starch and 0·60 lb protein for each gallon of milk produced. The minimum quantities of starch and protein required per cow per month are also given in Table I.

In order to obtain this starch and protein, a cow has to eat a certain amount of dry matter, but this must not exceed 33 lb a day. It was realised, however, that even if the total requirements of both starch and protein came from only one food, this maximum would still not be reached except in the case of mangolds. Mangold production proved too expensive to enter into the programme, and so it was

TABLE I

Monthly yield and nutrition requirements per dairy cow

	Milk Yield (gal)	Starch Required (lb)	Protein Required (lb)
January	71	394·5	65·85
February	68	366·0	61·80
March	75	404·5	68·25
April	75	397·5	67·50
May	82	422·0	72·45
June	70	385·0	64·50
July	62	372·0	60·45
August	56	357·0	56·85
September	56	350·0	56·10
October	57	359·5	57·45
November	62	365·0	59·70
December	66	382·0	62·85
Total	800	4555·0	753·75

unnecessary to consider the dry matter restriction in the remainder of the calculation. A cow needs a minimum quantity of bulky food in its diet, and it was decided that at least sufficient starch to supply maintenance requirements should come from food of this type, and should be grown on the farm.

The gross income of £128·5 per cow was calculated from the milk prices in force during the period October 1955-September 1956; the variable costs were estimated at £45 per cow for the first 40 cows, and £51 per cow thereafter, to allow for mortgage payments on the new buildings which would be required if the herd size exceeded 40. The net income was taken as the difference between these figures.

A representative selection of crops was chosen for possible inclusion in the optimum land utilisation programme. In the case of grazing, one acre yields a varying amount of food from April to mid-October, reaching a peak in May, and then falling off gradually. Hay and silage could also be grown; these crops could be harvested either once or twice during the summer and fed at any time during the year. If only one cut was taken, the land could then be grazed from August onwards; this method, known as aftermath grazing, is often used because the extra acreage supplements the main pasture when the output is falling off. Kale could be either cut and carted, or folded (grazed off the field). The former method gives a higher yield per acre, but has higher labour costs. Kale cannot be stored, but must be fed immediately it is harvested. Mangolds, on the other hand, must be stored after harvest, and for various technical reasons are only available for feeding from 1st January to 30th April. The foods mentioned so far—pasture, hay, silage, kale, and mangolds—are bulky foods, and must supply all the starch required for maintenance during the year. Oats and beans are harvested and may then be stored, so that they are available as fodder throughout the year. It was found that the year could be divided into eight periods, in each of which a different combination of crops or grazing output was available. The crops and the periods

TABLE II

Availability of crops

	Grazing	Aftermath Grazing	1 or 2 Cuts Hay or Silage	Cut and Carted:		Folded Kale	Mangolds	Oats or Beans
				Marrow-stem Kale	1000-head Kale			
Mid-October—December	O	O	I	I	O	I	O	I
January—March	O	O	I	O	I	I	I	I
April	I	O	I	O	O	I	I	I
May	I	Ø	I	O	O	O	O	I
June	I	O	I	O	O	O	O	I
July	I	O	I	O	O	O	O	I
August	I	I	I	O	O	O	O	I
September—Mid-October	I	I	I	I	O	I	O	I

I Available O Not available

in which they were available are shown in Table II. The yield of starch and protein, and the variable costs of production for each crop were calculated on an acreage basis.

The programme also allowed that oats, beans, and two typical manufactured foods—dairy cake and ground-nut cake—could be purchased to supplement home-grown crops.

Setting up the Original Matrix

Altogether 91 different crops were available; the quantity of each is denoted by x_j, in acres or tons according to whether the crop was grown or purchased. The number of cows with a net income of £83·5 is denoted by x_{92} ($\leqq 40$), and the number with a net income of £77·5 by x_{93}.

Twenty-seven inequalities, D_1—D_{27}, were imposed on the x_j in order to comply with the given restrictions.

D_1 ensured that the acreage did not exceed 100:

$$\sum_{1}^{91} h_j x_j \leqq 100 \quad \text{where } h_j = 1 \text{ for crops grown}$$

$$h_j = 0 \text{ for purchased foods}$$

D_2—D_{25} ensured that the cow's requirements in total starch, total protein, and starch for maintenance were met. For example, D_{20} concerns total starch in August. Suppose crop j supplies $1000a_j$ lb of starch for consumption in August. This will be zero for all crops supplying food for any period other than August. One cow requires 357 lb starch in August.

$$\sum_{1}^{91} a_j x_j \geqq 0{\cdot}357(x_{92} + x_{93})$$

or

$$-\sum_{1}^{91} a_j x_j + 0{\cdot}357(x_{92} + x_{93}) \leqq 0$$

The factor of 1000 was introduced to reduce all a_j to between 1 and 10 for ease of computation. Similarly a factor of 100 was introduced into the protein restrictions.

The inequality D_{26} restricted the acreage available for aftermath grazing, x_{66}, to the total acreage from which one cut of grass for hay or silage was taken:

$$x_{66} - \sum_{1}^{91} l_j x_j \leqq 0 \quad \text{where } l_j = 1 \text{ for 8 crops of single-cut hay and 8 crops of single-cut silage}$$

$$l_j = 0 \text{ otherwise}$$

The 27th restriction limited to 40 the number of cows which could be carried before new buildings were needed.

$$x_{92} \leqq 40$$

The general form of the ith inequality is denoted by

$$\sum_{j=1}^{93} a_{ij}x_j \leqq b_i$$

The income from the jth enterprise is denoted by c_j.

A portion of the original matrix, involving those crops available in August, is shown in Table III.

Method

As there were 27 restrictions and 93 possible enterprises, direct application of the Simplex method[1] of solution of a linear programming problem was quite impracticable without an electronic computer. However, a modification of this method, the Inverse Matrix method, was found to be suitable for a desk machine in this case. Twenty-seven non-negative variables x_{93+i} $(i=1, \ldots, 27)$ are introduced to transform the inequalities into equalities.
The ith inequality then becomes

$$\sum_{j=1}^{93} a_{ij}x_j + x_{93+i} = b_i$$

or

$$\sum_{j=1}^{120} a_{ij}x_j = b_i, \text{ where } a_{ij} = 0 \text{ for } j>93 \text{ except for } j = 93+i \text{ when it is 1.}$$

Initially the basic variables x_{93+i} are chosen to be equal to b_i, and the non-basic variables x_j are taken as zero. Now if it is known that the basic variables in the final solution are $x_{p_1}, x_{p_2}, \ldots, x_{p_{27}}$, where the x's may be chosen from either the original basic or non-basic variables, then the final solution can be computed directly. The 27×27 square matrix $\mathbf{K}=(a_{is})$ is inverted to give $\mathbf{K}^{-1}=(k_{is})$ where $s=p_1, p_2, \ldots, p_{27}$. (The p's may be chosen in any convenient order to simplify the matrix inversion.) The values of the constants a'_{ij}, b'_i, c'_j in the final programme are then given by the following equations:

$$\left.\begin{aligned} a'_{ij} &= k_{ip_1}a_{1j} + k_{ip_2}a_{2j} + \ldots + k_{ip_{27}}a_{27j} \\ b'_i &= k_{ip_1}b_1 + k_{ip_2}b_2 + \ldots + k_{ip_{27}}b_{27} \\ c'_j &= c_j - d_{p_1}a_{1j} - d_{p_2}a_{2j} - \ldots - d_{p_{27}}a_{27j} \end{aligned}\right\} \begin{matrix} i = 1, 2, \ldots, 27 \\ j = 1, 2, \ldots, 120 \end{matrix}$$

where $d_s = c_{p_1}k_{1s} + c_{p_2}k_{2s} + \ldots + c_{p_{27}}k_{27s}$

and where c_{p_i} is the income from the p_ith enterprise. In the following discussion, the primes are omitted for clarity; the non-primed constants a_{ij}, b_i, and c_j now refer to the postulated final programme.

This Inversion method can be used to compute several stages of the Simplex method at once, irrespective of whether the final solution gives the maximum income or not. In practice, however, it is only useful if one has a good idea which enterprises will be included in the final solution. In this problem, preliminary programmes using the ordinary Simplex method were carried out for one summer and one winter period. From these results, a final programme was postulated for the overall problem. The 27×27 matrix of these enterprises was then inverted by the method of Hayes and Fox.[2] The complete basis for the postulated solution was then calculated from the equations given above.

If the postulated solution derived as above does not give the maximum income, there are two possible ways in which adjustment may be required. One is that some of the b_i in the derived solution may be negative (thus showing a negative quantity of crops to be grown or bought, or that less than the requisite minimum quantity of starch or protein is produced). The second is that, as in the ordinary Simplex method, not all the c_j's in the derived solution are negative. If some particular $b_r < 0$, the option r must be removed from the programme and replaced by some t say, which satisfies if possible the following conditions:

(i) $a_{rt} < 0$

(ii) $\dfrac{b_r}{a_{rt}} = \min\left\{\dfrac{b_i}{a_{it}}\right\}$ for all $\dfrac{b_i}{a_{it}} > 0$

If no such t exists, b_r has to be increased in several stages by first bringing in an enterprise that satisfies (i) but not (ii), and then trying to find one that satisfies both (i) and (ii). If more than one such t exists, the one must be chosen such that:

(iii) $c_t > 0$

(iv) of those t satisfying (iii), $\dfrac{c_t}{a_{rt}}$ is a minimum

It was usually found easier to consider the corrections for $b_i < 0$ before those for $c_j > 0$.

In this particular problem, it was found that if some enterprise entered the basic variables in order to make either of the above alterations, then it would usually replace some other enterprise or restriction relating to the same time-period and that the exchange would have very little effect on coefficients relating to other time-periods. This made it easy to see the approximate effect of several changes at once, and hence the Matrix Inversion method could profitably be applied a second time. In most cases it was possible to reach the final solution with only two inversions.

TABLE III

Original matrix for August

	b_i	Grazing	After-math Grazing	2 Cuts Hay	1 Cut Hay	2 Cuts Silage	1 Cut Silage	Oats	Beans	Bought Oats	Bought Beans	Dairy Cake	Ground-Nut Cake	Cows	Cows
D_1—Land	100	1	0	1	1	1	1	1	1	0	0	0	0	0	0
D_{20}—Starch	0	−0·2557	−0·2557	−1·568	−1·098	−1·980	−1·383	−1·478	−1·035	−1·344	−1·478	−1·389	−1·635	0·357	0·357
D_{21}—Protein	0	−0·571	−0·571	−2·69	−1·88	−3·05	−2·13	−1·97	−3·14	−1·79	−4·48	−3·36	−9·18	0·5685	0·5685
D_{22}—Maintenance Starch	0	−0·2557	−0·2557	−1·568	−1·098	−1·980	−1·383	0	0	0	0	0	0	0·217	0·217
D_{26}—Land for Aftermath Grazing	0	0	1	0	−1	0	−1	0	0	0	0	0	0	0	0
D_{27}—Herd Size	40	0	0	0	0	0	0	0	0	0	0	0	0	1	0
c_j—Income	0	−4·5	0	−11·6	−9·4	−18·9	−14.2	−14·1	−17·6	−23.33	−32·33	−35·0	−42·0	83·5	77·5

Results

The solution to the original problem is given in Table IV; it was obtained after six alterations to the first estimate of the solution.

TABLE IV

Optimum programme

Crops:	Acres
Marrow-Stem Kale (September—mid-October)	4·0
Marrow-Stem Kale (mid-October—December)	12·2
Folded Kale (January—March)	25·2
Folded Kale (April)	7·0
1 Cut of Silage (August) and Aftermath Grazing	1·2
Grazing	50·4
	100·0
Purchased Foods:	
Oats purchased for consumption from May to August	30·0 tons
Ground-nut Cake purchased for consumption from September to April	15·6 tons
Herd Size	68·4 cows
Minimum restrictions on protein are not taken up from May to August. Minimum restrictions on starch for maintenance are not taken up from September to June.	

It was noticed that oats were bought during the summer months, as they are the cheapest source of starch available for purchase, and grazing is short of starch compared with protein relative to a cow's requirements. During this period, more than the minimum quantity of protein required by the cows was available. Similarly, kale supplies comparatively little protein relative to a cow's requirements, and ground-nut cake, the cheapest source of protein available for purchase, is fed from September to April. The restriction on starch for maintenance is only taken up during July and August. After subtracting £275 as a fixed cost for rent, the net income from the 68·4 cows per 100 acres is £2532.

The solution was then used to find optimum programmes for conditions pertaining on various other (hypothetical) farms. The first restriction imposed was that kale should not be folded after 31st December. Such grazing is often impracticable, particularly on clay land which tends to become waterlogged in winter. The two enterprises involved (folded kale in January to March and in April) were removed from the programme in the usual way by imagining a large increase, M, in their cost of production. It was immediately apparent that cut and carted thousand-headed kale (January to March) and single-cut silage (April) would become new basic variables. It could then be seen that these two changes would produce positive c_j's for oats purchased for feeding in these periods, and that these would enter at the expense of the starch in excess of that necessary for maintenance in the same period. Accordingly the matrix comprising these four changes (a 27×27 matrix where only the first four rows contained non-unit vectors) was inverted. It was found that as a result of these

changes, all $b_i > 0$, but that two of the $c_j > 0$ (those from maintenance starch for July and August). The complete columns were calculated for these two enterprises in order to find which enterprises they would replace in the basic solution. These turned out to be single-cut silage for feeding in August, and oats purchased for feeding in May, respectively. Accordingly a 27×27 matrix comprising these six changes from the original optimum was inverted. In this matrix the first six rows contained non-unit vectors, and it will be noted that the first four columns of the lower, and rows of the upper, triangular resolution matrices are the same as those of the first inversion. If the final array is required in full, this method is considerably shorter than calculating the complete array after only the first four changes, and then calculating the last two completely separately. In this case, the six changes gave all $b_i > 0$ and all $c_j < 0$, though the process could, of course, have been repeated. In this solution the number of cows fell to 62·6, the grazing acreage rose to 71·0 and the net profit fell to £2382.

Another programme, in which no kale at all could be grown after 31st December, was also calculated.

It will be seen that this method has a somewhat limited application, as it involves being able to predict fairly accurately which enterprises will be included in the final solution. It is hoped, however, that applications may be found in other fields.

Acknowledgements

I should like to thank the Farm Economics Branch of the School of Agriculture, University of Cambridge, where this work was carried out, for permission to publish this paper. I am grateful to Mr C. S. Barnard for suggesting the problem and supplying the agricultural data.

REFERENCES

[1] Vajda, S. (1955). 'An outline of linear programming', *J. R. Statist. Soc.*, B, **17**, 165.

[2] Fox, L. and Hayes, J. G. (1951). 'More practical methods for the inversion of matrices', *J. R. Statist. Soc.*, B, **13**, 83.

ON COMPARING TWO OBSERVED FREQUENCY COUNTS

MORTON KUPPERMAN
The George Washington University, Washington, DC

A simplified method of calculating the χ^2 statistic for deciding whether or not two sample frequency counts came from the same unspecified population is discussed. An alternative statistic is presented; it is based on information-theory considerations and has the advantage of being simpler to calculate.

Introduction

A problem that arises frequently in the applications of statistics to such diverse fields as psychology, medicine, biology, education, agriculture, etc., is to decide whether or not two observed frequency counts came from the same unspecified population. It is interesting to note that this problem was solved (for most practical purposes) in 1911 by K. Pearson.[1] The appropriate statistic (a χ^2 statistic) is well known to statisticians and may be found in many of the standard statistics textbooks. An equivalent simplified version of the statistic, more suitable for numerical calculation, does not seem to be as widely known by applied statisticians as it should be. It is the purpose of this article, after describing the simplified statistic and giving an example of its use, to introduce an alternative statistic based on information-theory considerations and having the advantage of ease of calculation, in addition to possessing certain desirable theoretical properties.

Let us suppose that we have two independent samples, consisting of N_1 and N_2 independent observations respectively, each sample having been randomly drawn from some c-valued population. The c-valued population may be either a discrete distribution of c classes or a continuous distribution that has been grouped into c categories. Let the observed frequencies for the first sample be given by $f_1, f_2, \ldots, f_c$, where c is the number of categories involved, and let the second sample's frequencies be $g_1, g_2, \ldots, g_c$. We have

$$\sum_{i=1}^{c} f_i = N_1 \text{ and } \sum_{i=1}^{c} g_i = N_2$$

Now write the observed frequencies in the following table:

	Frequency (first sample)	*Frequency (second sample)*	*Total*
	f_1	g_1	f_1+g_1
	f_2	g_2	f_2+g_2
	...	...	...
	f_c	g_c	f_c+g_c
Total	N_1	N_2	N_1+N_2

The χ^2 Statistic

The simplified version of the χ^2 statistic that we calculate is

$$\chi^2 = \frac{(N_1+N_2)^2}{N_1 N_2}\left(\sum_{i=1}^{c}\frac{f_i^2}{f_i+g_i} - \frac{N_1^2}{N_1+N_2}\right)$$

For large samples this statistic is distributed approximately like a theoretical χ^2 variate with $c-1$ degrees of freedom when both samples have in fact been randomly and independently drawn from the same population. This statistical test is due to Pearson, although the correct number of degrees of freedom for the asymptotic (large-sample) distribution was determined by R. A. Fisher.[2] The particular form of the statistic above, which is convenient for numerical computation, is occasionally attributed to A. E. Brandt and G. W. Snedecor, although Snedecor[3] attributes it to Snedecor and M. R. Irwin[4] (see reference 3, pp. 227-228). Readers interested in the mathematical theory underlying this statistical test will find a discussion and rigorous proofs in H. Cramér,[5] pp. 447-448, where the above formula may be found. Although our present concern is with testing the null hypothesis of homogeneity of two sample frequency counts, readers will recognise χ^2 above as being the same statistic used to test the null hypothesis of independence in a $c \times 2$ contingency table.

To illustrate the use of χ^2, let us use the numerical data given in Table 38, reference 6, pp. 90-91, representing two samples of Czech vowel phonemes. (Herdan's book,[6] incidentally, represents an interesting and important application of statistical methods and information theory to the study of language.) For this example $c = 12$. The following table shows the observed frequency counts:

Category	*Frequency (first sample)*	*Frequency (second sample)*	*Total*
1	160	150	310
2	137	142	279
3	106	125	231
4	74	89	163
5	35	39	74
6	29	30	59
7	28	35	63
8	29	41	70
9	19	22	41
10	6	11	17
11	8	11	19
12	13	4	17
Total	644	699	1343

We calculate χ^2 as follows. It does not matter which of the two samples is taken to give the f_i's of the formula since we shall find that χ^2 is also given (exactly) by

$$\chi^2 = \frac{(N_1+N_2)^2}{N_1 N_2}\left(\sum_{i=1}^{c}\frac{g_i^2}{f_i+g_i} - \frac{N_2^2}{N_1+N_2}\right)$$

Hence in practice we may take as the column to use in the numerator of the terms in parentheses the column with the generally smaller frequencies; in our example above we shall choose the first sample. Thus

$$\chi^2 = \frac{1343^2}{(644)\ (699)}\left(\frac{160^2}{310}+\frac{137^2}{279}+\frac{106^2}{231}+\frac{74^2}{163}+\frac{35^2}{74}+\frac{29^2}{59}\right.$$
$$\left.+\frac{28^2}{63}+\frac{29^2}{70}+\frac{19^2}{41}+\frac{6^2}{17}+\frac{8^2}{19}+\frac{13^2}{17}-\frac{644^2}{1343}\right)$$
$$= 11{\cdot}12$$

Herdan obtains 13·7, but this value may possibly be a typographical error for 11·7, which value one obtains by using the data in his Table 38. However, his expected values for the first category should have been rounded to 149 and 161 for samples 1 and 2 respectively instead of to 148 and 162. Using the corrected expected values, we obtain $\chi^2 = 11{\cdot}43$ following his procedure; the discrepancy between this figure and 11·12 is a consequence of rounding the expected values to the nearest unit. The number of degrees of freedom is 11, and on entering a table of percentage points of the χ^2 distribution we see that the probability of getting a value of χ^2 greater than 11·12 lies between 30 per cent and 50 per cent. If we assume beforehand a 5 per cent level of significance, then since the critical region for 11 degrees of freedom at a 5 per cent level of significance is the region with values of χ^2 greater than 19·68, we see that the two sample frequency counts may very reasonably be taken to have come from the same population. Thus the two observed samples may be taken to be homogeneous.

The Information Statistic

There exists another statistic, not well known to applied statisticians, that may be used to test this same null hypothesis of homogeneity. It is based on I(1:2), the Kullback-Leibler measure of information (see reference 7), and is used extensively in the book by S. Kullback[8] on the application of information theory to statistical analysis. The statistic to be considered here was previously described by S. S. Wilks[9] in his application of the likelihood-ratio test to contingency tables. Using the same notation as before, where the f_i's and the g_i's are the observed frequencies in the first and second sample respectively, we define the statistic (which I have elsewhere[10] termed 'information statistic') as

$$2\hat{I} = 2\left\{\sum_{i=1}^{c} f_i \log f_i + \sum_{i=1}^{c} g_i \log g_i + (N_1+N_2)\log(N_1+N_2)\right.$$
$$\left. - \sum_{i=1}^{c}(f_i+g_i)\log(f_i+g_i) - N_1 \log N_1 - N_2 \log N_2\right\}$$

where the logarithms are natural (or Napierian) logarithms (to the base e).

It may be proved that $2\hat{I}$, the information statistic, is asymptotically distributed as a theoretical χ^2 variate with $c-1$ degrees of freedom under the null hypothesis that both samples are from the same population. One proof makes use of the fact that $2\hat{I}$ is equal to $-2\log_e\lambda$, where λ is the likelihood-ratio criterion of J. Neyman and E. S. Pearson[11] for testing the null hypothesis that the two samples came from the same c-valued population with unknown probabilities.[9, 10, 12] Another proof makes use of a more general theorem on the asymptotic distribution of information statistics.[10] It may be noted that Herdan, in the aforementioned book,[6] has drawn attention to the close relationship between χ^2 and the entropy H (which is related to $\hat{I}$ in the present paper).

The information statistic $2\hat{I}$ is very simply calculated (much more easily than χ^2) since it involves only sums and differences of terms like $n\log_e n$, which are now available from a valuable 10-decimal-place table of $n\log_e n$ included as an appendix in Kullback's book.[8] The values of $n\log_e n$ (mentally rounded to 3 or 4 decimal places) are read out of the table directly into an adding machine; no squarings, multiplications, or divisions are involved, other than the factor 2 in $2\hat{I}$. Wilks[9] noted in 1935 that $2\hat{I}$ (which he terms $-2\log\lambda$, since he is concerned with the likelihood-ratio criterion) is simpler to calculate than χ^2, but this fact seems to have been overlooked by present-day writers of statistics textbooks, probably because of the nonavailability hitherto of tables of $n\log_e n$. However, A. M. Mood,[13] p. 276, does consider the use of the likelihood-ratio criterion λ in his treatment of contingency tables, and presents the problem of testing the homogeneity of several independent sample frequency counts in problem 39, p. 287 of reference 13. In a recently published article, I. J. Good[14] also points out the computational simplicity of $2\hat{I}$ (or, rather, of $-2\log\lambda$) and interestingly suggests the use of tables of $2n\log_e n$ rather than $n\log_e n$. In addition, he notes that 'the calculation is less "well-conditioned" than for χ^2, in the sense that more significant figures must be held'.

By way of illustration of the calculation of $2\hat{I}$, consider the example previously used. We find easily that

$$2\hat{I} = 2(2828{\cdot}609+3106{\cdot}260+9673{\cdot}174-6858{\cdot}942-4165{\cdot}198-4578{\cdot}206)$$
$$= 11{\cdot}39$$

and we arrive at the same conclusion as by using χ^2. It is suggested that the reader go through the actual steps of the preceding calculations of χ^2 and $2\hat{I}$ in order to see for himself the simplicity of the method when using the information statistic. For convenience a table of the pertinent values of $n\log_e n$ is appended. The reader should also note the close correspondence between the values of χ^2 and $2\hat{I}$, the difference, in this example, being only 0·27.

The table of $n\log_e n$ in reference 8 unfortunately does not go beyond

$n = 1000$ (presumably because of space limitations). It would be most useful to have the table extended to at least $n = 10\,000$, for in practice frequencies greater than 1000 often arise. A more useful table would be an extensive table of $2n\log_e n$, preferably rounded to 3 decimal places.

n	$n\log_e n$	n	$n\log_e n$
4	5·545	70	297·395
6	10·751	74	318·501
8	16·636	89	399·489
11	26·377	106	494·325
13	33·344	125	603·539
17	48·165	137	674·037
19	55·944	142	703·727
22	68·003	150	751·595
28	93·302	160	812·028
29	97·652	163	830·281
30	102·036	231	1257·198
35	124·437	279	1571·108
39	142·879	310	1778·337
41	152·256	644	4165·198
59	240·575	699	4578·206
63	261·017	1343	9673·174

Acknowledgement

I should like to thank Dr G. Herdan for permission to use Table 38 from his book *Language as Choice and Chance*.

Note Added in Proof

Since the above paper was submitted for publication, my attention has been drawn to a paper by B. Woolf,[15] who applies the logarithm of the likelihood-ratio statistic to the analysis of contingency tables and presents a valuable table of $2n\log_e n$, to 4 decimal places, for n up to 2009.

REFERENCES

[1] Pearson, K. (1911). 'On the probability that two independent distributions of frequency are really samples from the same population', *Biometrika*, **8**, 250.

[2] Fisher, R. A. (1922). 'On the interpretation of χ^2 from contingency tables, and the calculation of P', *J. R. Statist. Soc.*, **85**, 87.

[3] Snedecor, G. W. (1956). *Statistical Methods Applied to Experiments in Agriculture and Biology*, 5th edition. The Iowa State College Press, Ames.

[4] Snedecor, G. W. and Irwin, M. R. (1933). 'On the chi-square test for homogeneity', *Iowa State College Journal of Science*, **8**, 75.

[5] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton.

[6] Herdan, G. (1956). *Language as Choice and Chance*. P. Noordhoff N. V., Groningen.

[7] KULLBACK, S. and LEIBLER, R. A. (1951). 'On information and sufficiency', *Ann. Math. Statist.*, **22**, 79.

[8] KULLBACK, S. (1959). *Information Theory and Statistics.* John Wiley, New York.

[9] WILKS, S. S. (1935). 'The likelihood test of independence in contingency tables', *Ann. Math. Statist.*, **6**, 190.

[10] KUPPERMAN, M. (1957). *Further Applications of Information Theory to Multivariate Analysis and Statistical Inference.* Unpublished Ph.D. dissertation, The George Washington University.

[11] NEYMAN, J. and PEARSON, E. S. (1928). 'On the use and interpretation of certain test criteria for purposes of statistical inference', *Biometrika*, **20A**, 175 and 263.

[12] WILKS, S. S. (1938). 'The large-sample distribution of the likelihood ratio for testing composite hypotheses', *Ann. Math. Statist.*, **9**, 60.

[13] MOOD, A. M. (1950). *Introduction to the Theory of Statistics.* McGraw-Hill, New York.

[14] GOOD, I. J. (1957). 'Saddle-point methods for the multinomial distribution', *Ann. Math. Statist.*, **28**, 861.

[15] WOOLF, B. (1957). 'The log likelihood ratio test (the *G*-test). Methods and tables for tests of heterogeneity in contingency tables', *Ann. Hum. Genet., Lond.*, **21**, 397.

ANALYSIS OF VARIANCE OF A LATIN SQUARE DESIGN WITH MISSING OBSERVATIONS*

CLYDE YOUNG KRAMER and SUZANNE GLASS†
Virginia Agricultural Experiment Station of the Virginia Polytechnic Institute

The authors here describe a general method for estimating several missing values and analysing the results so obtained. They give explicit formulae for all cases when two values are missing, formulae for some other specific cases, and an illustrative example.

Introduction

The statistical analysis of a Latin square design is sometimes complicated by the fact that some of the observations are missing. The usual method of analysis requires that the design be balanced, but missing values destroy this property of balance.

Allan and Wishart[1] were among the first writers to furnish a method of analysis for a single missing plot. They obtained a formula for estimating the missing observation by minimising the error sum of squares. The augmented data were then analysed in the usual way except that the numbers of degrees of freedom for the error and for the total sum of squares were decreased by one. Yates[4] showed that this method of analysis resulted in an approximate F-test for the treatment mean square, and he furnished a method for estimating the bias.

Much of the writing on missing plots in Latin squares has been done by Yates[5] and by Yates and Hale.[6] Articles by these authors give a method of analysis of Latin squares when a single observation is missing; when one or more rows, columns, or treatments are missing; and when one column and one or more single observations are missing. De Lury[2] brought a number of these methods together.

Yates[4] has shown how the method for dealing with one missing plot may be extended to the case where several observations are missing. However, an iterative procedure is necessary when this method is used to estimate more than one missing value.

This paper will deal with the estimation of several missing values in the Latin square design, by minimising the error sum of squares. Explicit formulae for each value missing will be derived for many special cases and a procedure will be given for the completely general

* A condensation of a thesis by S. Glass written under the direction of C. Y. Kramer and submitted to the Virginia Polytechnic Institute in partial fulfilment of the requirements for the M.S. degree in Statistics.

† Now at Roanoke College, Salem, Virginia.

case. A direct method of analysis of variance, not requiring a correction for bias in the treatment sum of squares, will be demonstrated by using the work of Glenn and Kramer.[3]

Model and Notation

The basis of the usual analysis of a Latin square design is the assumption of the model

$$y_{ij(m)} = \mu + \rho_i + \gamma_j + \tau_m + \epsilon_{ij(m)};\ i, j, \text{ and } m = 1, \ldots, k \quad \ldots . \quad (2.1)$$

where μ, ρ_i, γ_j, and τ_m are constants such that $\Sigma\rho_i = \Sigma\gamma_j = \Sigma\tau_m = 0$; μ, ρ_i, γ_j, and τ_m are the grand mean, row effects, column effects, and treatment effects, respectively. The $\epsilon_{ij(m)}$, representing random error effect, are independently and normally distributed with mean 0 and variance σ^2. The subscript in parentheses is not an independent subscript. Once values for i and j are chosen, the value of m is determined for a given Latin square arrangement. This notation will be used throughout this paper. In some instances, two subscripts will appear in parentheses, meaning that both of these subscripts depend upon the subscript outside the parentheses.

We shall denote row totals, column totals, and treatment totals for the remaining plots in any row, column, and treatment in which the missing plots appear by R'_i, C'_j, and T'_m; and shall denote the grand total of the observed values by G'. The term 'plots', derived from agricultural experimentation, is here used in an extended sense to denote the individual units to which the treatments are applied.

Estimates of Missing Values

Case of n Missing Values, No Two in the Same Row, Column or Treatment. Let the missing values be represented by $x_{fg(h)}$, $x_{rs(t)}$, $x_{uv(w)}$, . . ., n in all. By partially differentiating the expression for the error sum of squares with respect to the x's and equating the derivatives to zero the following system of equations is obtained:

$$\begin{aligned}
(k-1)(k-2)x_{fg(h)} + 2x_{rs(t)} + \ldots + 2x_{uv(w)} &= k(R'_f + C'_g + T'_h) - 2G' \\
2x_{fg(h)} + (k-1)(k-2)x_{rs(t)} + \ldots + 2x_{uv(w)} &= k(R'_r + C'_s + T'_t) - 2G' \\
\cdot \qquad \cdot \qquad \cdot \qquad \cdot \qquad &= \qquad \cdot \\
2x_{fg(h)} + 2x_{rs(t)} + \ldots + (k-1)(k-2)x_{uv(w)} & \\
= k(R'_u + C'_v + T'_w) - 2G' & \quad \ldots . \quad (3.1)
\end{aligned}$$

Defining the quantity P as follows:

$$P_{ij(m)} = k(R'_i + C'_j + T'_m) - 2G' \quad \ldots . \quad (3.2)$$

one may express the above system of equations in matrix notation as

$$\begin{bmatrix} (k-1)(k-2) & 2 & . & 2 \\ 2 & (k-1)(k-2) & . & 2 \\ . & . & . & . \\ 2 & 2 & . & (k-1)(k-2) \end{bmatrix} \begin{bmatrix} x_{fg(h)} \\ x_{rs(t)} \\ . \\ x_{uv(w)} \end{bmatrix} = \begin{bmatrix} P_{fg(h)} \\ P_{rs(t)} \\ . \\ P_{uv(w)} \end{bmatrix}$$

. . . . (3.3)

The solution of (3.3) is given by

$$x_{cd(e)} = \frac{[(2n-4)+(k-1)(k-2)]P_{cd(e)} - 2\sum_{\substack{\text{missing plots} \\ i \neq c}} P_{i(j)(m)}}{[(k-1)(k-2)+(2n-2)]\,[(k-1)(k-2)-2]} \quad \ldots . \quad (3.4)$$

where c, d, e take on successively the values corresponding to missing plots. The summation in the numerator refers to the sum of P values for all missing plots except $cd(e)$.

Case of n Missing Values, All in a Single Row, Column, or Treatment. Consider first the case in which n values are missing in the ith row and let them be designated by $x_{ig(h)}$, $x_{is(t)}$, $x_{iv(w)}$, ... As in the previous section a system of equations is obtained by partially differentiating the expression for the error sum of squares. This system is expressed in matrix notation as

$$\begin{bmatrix} (k-1)(k-2) & (2-k) & . & (2-k) \\ (2-k) & (k-1)(k-2) & . & (2-k) \\ . & . & . & . \\ (2-k) & (2-k) & . & (k-1)(k-2) \end{bmatrix} \begin{bmatrix} x_{ig(h)} \\ x_{is(t)} \\ . \\ x_{iv(w)} \end{bmatrix} = \begin{bmatrix} P_{ig(h)} \\ P_{is(t)} \\ . \\ P_{iv(w)} \end{bmatrix}$$

. . . . (3.5)

The formula for the estimation of n missing values, all in row i, is

$$x_{id(e)} = \frac{(k-n+1)P_{id(e)} + \sum_{\substack{\text{missing plots} \\ j \neq d}} P_{(i)j(m)}}{k(k-2)(k-n)} \quad \ldots . \quad (3.6)$$

where d, e take on successively the values corresponding to missing plots in row i; the summation of P values in the numerator refers to all missing plots in row i except $id(e)$.

In the same manner one obtains the formula for the estimation of n missing values, all in column j, as

$$x_{cj(e)} = \frac{(k-n+1)\,P_{cj(e)} + \sum_{\substack{\text{missing plots} \\ i \neq c}} P_{i(j)(m)}}{k(k-2)(k-n)} \quad \ldots . \quad (3.7)$$

and the formula for the estimation of n missing values, all in treatment m, as

$$x_{cd(m)} = \frac{(k-n+1)\,P_{cd(m)} + \sum_{\substack{\text{missing plots} \\ i \neq c}} P_{i(j)(m)}}{k(k-2)(k-n)} \quad \ldots . \quad (3.8)$$

Estimation of Two Missing Values. If only two values are missing, they must necessarily fall into one of the four categories discussed in the two previous sections. The appropriate formulae are obtained from (3.4), (3.6), (3.7), and (3.8).

If two values are missing, not in the same row, column, or treatment, let them be represented by $x_{fg(h)}$ and $x_{rs(t)}$; then

$$x_{fg(h)} = \frac{(k-1)(k-2)P_{fg(h)} - 2P_{rs(t)}}{[(k-1)(k-2)+2][(k-1)(k-2)-2]}$$
$$x_{rs(t)} = \frac{(k-1)(k-2)P_{rs(t)} - 2P_{fg(h)}}{[(k-1)(k-2)+2][(k-1)(k-2)-2]} \quad \ldots . \quad (3.9)$$

If two values are missing, both in row i, let them be represented by $x_{ig(h)}$ and $x_{is(t)}$. Then (3.6) gives

$$x_{ig(h)} = \frac{(k-1)P_{ig(h)} + P_{is(t)}}{k(k-2)^2} \qquad x_{is(t)} = \frac{(k-1)P_{is(t)} + P_{ig(h)}}{k(k-2)^2} \quad \ldots . \quad (3.10)$$

From (3.7) we find two missing values in the same column j to be

$$x_{cj(e)} = \frac{(k-1)P_{cj(e)} + P_{ij(m)}}{k(k-2)^2} \qquad x_{ij(m)} = \frac{(k-1)P_{ij(m)} + P_{cj(e)}}{k(k-2)^2} \quad \ldots . \quad (3.11)$$

If two values are missing, both in treatment m, then (3.8) gives

$$x_{cd(m)} = \frac{(k-1)P_{cd(m)} + P_{ij(m)}}{k(k-2)^2} \qquad x_{ij(m)} = \frac{(k-1)P_{ij(m)} + P_{cd(m)}}{k(k-2)^2} \quad \ldots . \quad (3.12)$$

General Case. When n missing values in a Latin square design are distributed arbitrarily over the rows, columns, and treatments, it is not possible to give an explicit equation for each value missing, covering all permissible values of n. However, it is possible to set down certain features which apply in general.

When the criterion of minimising the error sum of squares is applied, the unknowns (x's) representing missing values are obtained as the solutions of a system of linear equations which may be expressed in matrix notation as

$$\mathbf{WX} = \mathbf{P} \quad \ldots . \quad (3.13)$$

where $\mathbf{W}$ is an $n \times n$ symmetric matrix to be described below, $\mathbf{X}$ is an $n \times 1$ column vector of unknowns representing missing values, and $\mathbf{P}$ is an $n \times 1$ column vector of corresponding P values as defined by (3.2). In the cases considered above the solution was effected by inverting the $\mathbf{W}$ matrix to give a result of the form

$$\mathbf{X} = \mathbf{W}^{-1}\mathbf{P} \quad \ldots . \quad (3.14)$$

The particular structure of the $\mathbf{W}$ matrix depends on the manner in

which the missing values are distributed relative to one another over the rows, columns, and treatments. A complete description of the $n \times n$ matrix **W** is possible by numbering the vacant plots in any order from 1 to n and letting the ith missing value be x_i. The nature of the equations from which it results indicates that

(i) every diagonal element is $(k-1)(k-2)$;
(ii) if x_i and x_j are not in the same row, column, or treatment, element ij is 2;
(iii) if x_i and x_j are in the same row, column, or treatment, element ij is $(2-k)$.

Having written the column vector **X**, one may construct the matrix **W** from the above properties by referring to the data table. The column vector **P**, consisting of elements determined from (3.2) for each vacant plot, must be written in the order which corresponds to that of the x_i's. The system (3.13) is thus obtained.

Analysis of Variance

A Direct Method of Analysis. This method avoids the direct calculation of the bias in the treatment sum of squares arising from analysis of the augmented data (i.e., including estimates of missing values) by the standard method for a Latin square design. The estimates are first entered in the table. The analysis of the augmented data then proceeds as usual. However, in recording the sums of squares for total (before correction for the mean) a separation into two parts is made for later reference; one part corresponds to values originally present, the other part to values inserted. The approximate test of significance of the treatment effect is carried out on the augmented data, the degrees of freedom for total and error each being reduced by one for each missing value estimated. Since the bias always enhances the apparent significance of the treatment effect, one may stop if this test does not indicate significance.

If however, the approximate test indicates significance, one proceeds to find the correct treatment sum of squares to be tested against error. This is accomplished by considering the original data as a two-way classification. Missing values are then inserted in the two-way classification by the method described by Glenn and Kramer.[3] The usual sums of squares are then obtained, taking advantage of the partition of the total sum of squares that has been made. If we denote the error sum of squares obtained from the Latin square analysis as E and the error sum of squares obtained from the two-way classification as E', then the correct treatment sum of squares is found by subtracting E from E'.

Example

Illustrations, given below, of the application of the techniques described above are based on a Latin square experiment with two

missing values. The experiment was concerned with the application of commercial fertilisers to the Dunlap variety of strawberries. The yields are presented in Table I.

TABLE I

Yields in quarts

		Column 1	2	3	4	5	Original Totals	Latin Square Totals	Two-Way Totals
Row	1	A 31	D 58	C 56	B 44	E 40	229	229	229
	2	B 39	C 51	E 48	D 54	A 35	227	227	227
	3	C 51	B 48	A *51* *53*	E 74	D 53	226	277	279
	4	E 40	A *45* *47*	D 65	C 52	B 51	208	253	255
	5	D 55	E 53	B 41	A 46	C 47	242	242	242
Original Totals		216	210	210	270	226	1132		
Latin Square Totals		216	255	261	270	226		1228	
Two-Way Totals		216	257	263	270	226			1232

	A	B	C	D	E
Original Totals	112	223	257	285	255
Latin Square Totals	208	223	257	285	255

The values italicised and underlined in Table I have been estimated. The values in the upper left-hand corner are the estimates from the Latin square, the values in the lower right-hand corner are the estimates from the two-way classification. To obtain the Latin square estimates we first substitute the appropriate original totals in (3.2) to obtain

$$P_{42(A)} = 5(208+210+112) - 2(1132) = 386$$

and

$$P_{33(A)} = 5(226+210+112) - 2(1132) = 476$$

Then from (3.12)

$$x_{42(A)} = \frac{4(386)+476}{45} = 45 \qquad x_{33(A)} = \frac{4(476)+386}{45} = 51$$

After the estimated values are inserted in the data table the Latin square totals are obtained.

Next we compute the following sums of squares

$$\begin{aligned} \text{Total} &= 57\,724 + 4\,626 - 60\,319{\cdot}36 = 2\,030{\cdot}64 \\ \text{Row} &= 60\,654{\cdot}40 - 60\,319{\cdot}36 = 335{\cdot}04 \\ \text{Column} &= 60\,755{\cdot}60 - 60\,319{\cdot}36 = 436{\cdot}24 \\ \text{Treatment} &= 61\,058{\cdot}40 - 60\,319{\cdot}36 = 739{\cdot}04 \\ E &= 2\,030{\cdot}64 - 335{\cdot}04 - 436{\cdot}24 - 739{\cdot}04 = 520{\cdot}32 \end{aligned}$$

The approximate test is set up as shown in Table II. Since two values have been inserted, the degrees of freedom for total and error have each been reduced by two.

Since the approximate test suggests significance at the five per cent level, we proceed to find the correct treatment sum of squares to be tested against error. To accomplish this we first estimate the missing values for a two-way classification by the method given by Glenn and Kramer.[3] Using their equation (3.2) we obtain

$$\mathcal{Z}_{42} = 5(208+210) - 1132 = 958 \qquad \mathcal{Z}_{33} = 5(226+210) - 1132 = 1048$$

TABLE II

Approximate analysis of variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Rows	4	335·04	83·76	
Columns	4	436·24	109·06	
Treatments	4	739·04	184·76	3·55*
Error	10	520·32	52·03	
Total	22	2 030·64		

* Significant at 5 per cent level in approximate test.

and then from their equation (3.8) we obtain

$$x_{42} = \frac{16(958) - 1048}{255} = 47 \qquad x_{33} = \frac{16(1048) - 958}{255} = 53$$

After the estimated values are inserted in the data table the two-way classification totals are obtained and the usual sums of squares are computed (the 'Total' value utilises the 57 724 previously recorded separately).

$$\begin{aligned} \text{Total} &= 57\,724 + 5018 - 60\,712{\cdot}96 = 2\,029{\cdot}04 \\ \text{Row} &= 61\,080{\cdot}00 - 60\,712{\cdot}96 = 367{\cdot}04 \\ \text{Column} &= 61\,160{\cdot}00 - 60\,712{\cdot}96 = 457{\cdot}04 \\ E' &= 2\,029{\cdot}04 - 367{\cdot}04 - 457{\cdot}04 = 1\,204{\cdot}96 \end{aligned}$$

The correct treatment sum of squares is then found to be

$$E' - E = 1\,204{\cdot}96 - 520{\cdot}32 = 684{\cdot}64$$

The exact F ratio is then found to be

$$F = \frac{684{\cdot}64}{4(52{\cdot}03)} = \frac{171{\cdot}16}{52{\cdot}03} = 3{\cdot}29$$

The appropriate five per cent level of F is 3·48, and so the use of the exact test has brought the level of significance of the treatment effects below five per cent. At such levels inferences are essentially uncertain, and correction for the bias has increased the degree of uncertainty regarding the reality of the treatment effects.

REFERENCES

[1] Allan, F. E. and Wishart, J. (1930). 'A method of estimating the yield of a missing plot in field experimental work', *J. Agric. Sci.*, **20**, 399.

[2] De Lury, D. B. (1946). 'The analysis of Latin squares when some observations are missing', *J. Amer. Statist. Assoc.*, **41**, 370.

[3] Glenn, W. A. and Kramer, C. Y. (1958). 'Analysis of variance of a randomised block design with missing observations', *Applied Statistics*, **7**, 173.

[4] Yates, F. (1933). 'The analysis of replicated experiments when the field results are incomplete', *Emp. J. Exp. Agric.* **1**, 129.

[5] Yates, F. (1936). 'Incomplete Latin squares', *J. Agric. Sci.*, **26**, 301.

[6] Yates, F. and Hale, R. W. (1939). 'The analysis of Latin squares when two or more rows, columns, or treatments are missing', *Suppl. J. R. Statist. Soc.*, **6**, 67.

NOTE ON BRITISH UNEMPLOYMENT STATISTICS

DONALD BAILEY
Unemployment Insurance Section, Dominion Bureau of Statistics, Canada

In this article a Canadian statistician attempts an assessment of the extent to which British unemployment statistics represent the situation truly.

Unemployment can be defined in a variety of ways and its definition has been the subject of considerable controversy, mainly in the late 1940's. It is not intended to recapitulate that controversy here, except to restate the proposition that the amount of unemployment prevailing at any time will fluctuate not only with conditions of the labour market but with the manner in which unemployment is defined.

Persons with a recent and regular attachment to the labour force, and who suffer a temporary interruption of employment, give little trouble in measuring unemployment. There are, however, broadly speaking three groups of people who by reason of tenuous attachment to employment or questionable availability create problems of measurement. These are as follows:

1. Fringe groups who given favourable demand conditions are willing to work but who are not generally registered for employment at employment exchanges.
2. Groups who are in the count of unemployed because of administrative arrangements but whose labour force attachment may for various reasons be relatively weak.
3. Persons working less than full time who are sometimes categorised as under-employed.

Fringe groups, which are for the most part not represented in the official statistics, tend to be married women, widows, and pensioners, who are not regularly attached to the labour force but who may be attracted into the labour market in times of full employment. In the case of married women and pensioners willingness to take employment usually involves reversal of a previous decision to retire from the labour force.

In the British scheme of social insurance, as in any other, there are some persons who, while registered for employment and in general able to qualify for either Unemployment Benefit or National Assistance, have a tenuous attachment to the labour force. In the analysis of unemployment it is helpful to be able to identify such persons and study their characteristics.

Persons working reduced hours are in some circumstances under-

employed. It is sometimes contended that such partial unemployment should be converted to units of full-time unemployment and the official count of unemployment adjusted accordingly.

Official British statistics of unemployment are derived from a count of persons registering for employment at offices of the Ministry of Labour. Since application is a prerequisite for obtaining Unemployment Benefit or National Assistance (nearly all applicants are also claimants for either Insurance Benefit or Assistance) the adequacy of the count as a measure of unemployment depends in large part on the comprehensiveness of the National Insurance Scheme.

The British scheme of unemployment insurance covers practically all persons of working age (between the ages of 15 and 70 for men and 15 and 65 for women) employed under a contract of service. The chief exceptions are married women and widows, who may avail themselves of the option not to be insured for Unemployment Benefit.

The basis for permitting married women to be excepted from coverage is that they are already covered for certain benefits, including retirement pension, by virtue of their husband's insurance. Similarly, a widow entitled to pension or widowed mother's allowance through a deceased husband's insurance is not required to contribute to national insurance while working, because her pension becomes payable again as soon as she stops working. Having this in mind, such persons have to weigh the advantages of being covered for Unemployment Benefit against the extra cost involved.

It is estimated that there are some 2 260 000 married females who have 'opted out'.[1] This represents about 65 per cent of all occupied married females and therefore constitutes a potentially significant exception from coverage.

If we assume that the rate of unemployment for married females who have opted out is about 3 per cent, it appears that with total unemployment around 530 000 (December 1958), about 68 000 persons in this category might be without employment. (The rate for all women was about 2 per cent in December 1958; however, it seems probable that layoffs among married women with working husbands would be more numerous than for women generally.) It is known that a number of married women register for employment; however, it seems likely that the number who fail to register would not be less than 50 000. Such persons must be assumed to feel that, although wishing to find employment, they have nothing to gain from registration or to have decided to retire from active participation in the labour force.

Specific information regarding the number of widows in the labour force and the number who avail themselves of the opportunity to opt out of insurance is not available. From the 1 per cent sample of National Insurance Cards, however, the estimate of the number of widows and divorced women combined is about 450 000. Of these about 250,000,[1] or roughly 55 per cent, opt out. Since opportunity to opt out is not available to divorced women, the percentage for

widows is probably about the same as for married women, i.e. 65 per cent.

In the circumstances of December 1958, when the general level of unemployment for women was 2 per cent, about 9 000 widows and divorced women might be without employment. (This assumes that the level of unemployment for widows might be lower than the general level because of the probable tendency of employers to lay off women in this category more slowly than married women or spinsters.) Of these, perhaps 5 000 would not have the incentive provided by Unemployment Benefit to register.

Pensioners, who in most cases are not covered for Unemployment Benefit, constitute a fringe group which seems likely to be of considerable importance in times of full employment. There are at present over 5 300 000 retirement pensioners. Specific information regarding the number earning, i.e. having some labour force attachment, is not available although it is thought to be something like 500 000.

Of some 1 300 000 pensioners to whom the earnings rule would apply if they had earnings (men under the age of 70 and women under age 65), less than 3 per cent (about 40 000) have been found from special inquiries to have earnings in excess of the £3 limit, i.e. were engaged in employment to any considerable extent. From this it might be assumed that perhaps 2 per cent (or 80 000) of the 4 000 000 over the age where the earnings rule applies would have employment yielding £3 or more, making a total of 120 000 in this category. In the circumstances of December 1958, with the general index of unemployment at 2·4 per cent, some 25 000 to 30 000 of the group of pensioners defined as having an appreciable attachment to employment might lose their jobs; considering the low earnings limit used, 25 000 is probably the more reasonable figure. About 10 000 men and women of pensionable age were registered for employment at that time but doubtless a good many of these would be persons who had deferred their pensions. Hence the number of pensioners who do not register on becoming unemployed might perhaps be about 20 000.

On the basis of the above very rough estimates it appears that in the circumstances of December 1958 there might have been as many as 75 000 persons in the married women, widow, and pensioner category who would not have the incentive provided by Unemployment Benefit, and therefore might not register at Employment Exchanges on loss of employment. How many of these should be counted as unemployed and added to the official measure of unemployment depends on the degree of labour force attachment, or more specifically the intensity of desire for employment, which one attributes to such fringe groups.

In the case of working widows it seems that labour force attachment is not likely to be much different from that for single women. Thus if they opt out of insurance, in most cases it is likely to be due to the cost involved (weighed against being able to fall back on their pension in

the event of unemployment) rather than to the transitory character of the attachment. If this point of view is accepted it would seem likely that on being laid off most widows would register for employment at Employment Exchanges as non-claimants. There will however be a minority, particularly in one-industry towns, who will probably not register because of the belief that no work is available.

The attachment of married women who fail to register for employment on separation from their jobs is even more difficult to assess than for widows. It has been suggested by H. A. Turner that a high proportion of this group are anxious to find work but do not register because of lack of incentive.[2] The same writer contends that a secondary effect of the laying off of married women is to discourage the re-entry into employment of other married women.[3]

In deciding how to categorise pensioners who have lost their jobs one is again faced with the difficult problem of determining labour force attachment. It has been suggested by Professor Titmuss[4] that the sharp rise in living costs of recent years has exerted pressure on many pensioners to seek employment. This might be taken to imply that a high proportion of those who have lost their jobs are actively searching for other employment, and that there are in addition many pensioners who, although not actively seeking work, would be happy to have a job if given the opportunity of suitable employment.

An alternative point of view to that cited above is that in many instances when married women and pensioners take jobs a previous decision to retire is involved. Of such persons Professor Slichter has remarked: 'The decision for many persons to work or not to work is a close one. The location of a new plant with attractive working conditions and good wages will cause many persons to shift from the decision not to work to a decision to accept work. In many cases the right sort of part time job with convenient hours will be gladly accepted but not sought for.'[5]

Where persons are attracted into employment in the circumstances mentioned by Professor Slichter, it seems unlikely that they would have the same interest in employment when these specially favourable conditions no longer exist; for example, if employment was only available at lower wages or less convenient hours, or involved a considerable amount of travel. It is then questionable if any considerable proportion of such persons should be considered as unemployed if not registered for employment.

When married women have traditionally formed a considerable proportion of the regular working force of a local industry the situation differs from that sketched above. As H. A. Turner[6] suggests, the intensity of interest on the part of such women in finding employment is probably not much different from that of single women losing their jobs in the same industry. There is no information on the number of such persons; however, it still appears likely that most married women and pensioners not covered for Unemployment Benefit who are laid

off and fail to register have resumed their previous status of not being in the labour force.

With regard to the contention that there are many persons with no recent employment history interested in finding jobs in addition to those registered for employment, the results of a Special Survey some years ago by the United States Bureau of the Census is instructive. This Survey[7] revealed that some 3 000 000 persons classified as 'not in the labour force', on the basis of lack of activity in seeking employment during the survey week, indicated that they would 'like to have a job'. It should be noted that of the 3 000 000 about 560 000 stated that they would have looked for work except that they believed no work was available, were temporarily ill or expected to return to a job.

To raise the question of the extent to which persons who are not registered for employment, whether or not they are actively seeking work, are in fact unemployed, is to raise the question of how full is full employment. As Professor Reynolds[8] has noted, the goal of full employment tends to recede as it is approached. Presumably almost everyone who is in any degree employable will accept employment under certain optimum conditions. Having in mind the moral overtones of expressing willingness to take employment, it would seem unrealistic, however, to suggest that any full employment policy would ever undertake to provide jobs for everyone who under any circumstances might take employment. Therefore willingness to register at an Employment Exchange would appear to provide a reasonable, if not perfect, basis for distinguishing between those whose situation constitutes a problem requiring action to facilitate placement in employment and those whose situation does not.

Unemployed persons who register for employment at an Employment Exchange are required to report twice a week (in most countries weekly and even bi-weekly reporting is common), thus ensuring ample exposure to employment opportunities. If it is suspected that a claimant is in fact employed or is malingering he may be required to report even more frequently. This also ensures a minimum lag in adjusting the Register to take account of persons who find jobs by their own efforts between reporting times. Thus the count of the Register at any time tends to give a representative picture of total unemployment at the time the count is taken.

Although official statistics provide a reasonably acceptable measure of total unemployment, the figure includes some persons who have a much weaker labour force attachment than others and it would be useful for the researcher into unemployment problems to be able to identify such cases. For this purpose the provision of certain additional data is suggested.

Over recent years between 25 and 30 per cent of all registered unemployed persons have been in receipt of National Assistance. The majority of the groups receiving Assistance are not also receiving

Unemployment Benefit, although there appears to be some positive correlation between the proportion of cases involving supplementation and the level of unemployment.

In times of full employment, officially defined as unemployment not exceeding 3 per cent, Assistance cases, in addition to forming a substantial percentage of those registered for employment, tend to include many persons who have a difficult time re-establishing themselves in employment. It would therefore be useful to segregate Assistance cases in the monthly statistics of persons registered for employment.

Within the group receiving Assistance a further sub-division into short-term and long-term cases would be desirable. Broadly speaking, the short-term groups would consist mainly of those temporarily unable to qualify for Unemployment Benefit because of recent arrival in the country, and claimants under suspension, disallowance, disqualification or not yet rated. Long-term cases would consist mainly of those whose Unemployment Benefit rights were exhausted and who were unable to requalify. Both short- and long-term categories would include some persons receiving Assistance as supplementation of Insurance Benefits.

It is appreciated that the suggested division into short- and long-term groups might involve arbitrary decisions in certain borderline cases. But as a rough distinction between the two groups it would appear, for example, that under current conditions persons in receipt of Assistance for six months or more might reasonably be classified as long-term cases.

Something like two-thirds of Assistance cases appear likely to fall into the long-term category[9] and it is the size and composition of this group that is of special interest to the person analysing unemployment statistics. The majority of such persons are reported by the National Assistance Board as having been without work for periods of such duration as to suggest great difficulty in their being reabsorbed into regular employment.

A special inquiry conducted by the National Assistance Board in 1956 throws considerable light on the labour force attachment of long-term Assistance cases. At the time of the inquiry, Assistance cases numbering about 56 000 comprised approximately 25 per cent of all applicants for employment. Of these about 14 000 ceased receiving Assistance before being interviewed or were excluded because they had been on Assistance for less than two months. An additional 9000 over the age of 60 were also excluded. The circumstances of 32 000 remaining were investigated in the course of the inquiry.

The inquiry found that 40 per cent had done no work in the two-year period prior to interview and only about one-third had worked for as much as six months. Further, about 72 per cent of those interviewed had a physical or mental handicap which to some degree limited their chances of being permanently reabsorbed into employment.

Although no further inquiry was undertaken in 1957, it is instructive to encounter the following statement in the Board's report for that year:

'In conditions of full employment a high proportion of unemployed persons receiving assistance, although considered fit enough to be required to register for work, are in fact suffering from disabilities which greatly reduce their chances of securing regular employment.'[10]

The official percentage of unemployment rose from about 1·5 per cent in 1956 and 1957 to 2·4 per cent in 1958.[11] This undoubtedly had the effect of introducing into the long-term Assistance category a considerable number of persons immobilised in areas affected by structural unemployment; total Assistance cases, for example, rose from 85 000 in December 1957 to 150 000 in December 1958. Bearing in mind, however, that protracted unemployment produces physical and psychological effects which reduce employability in the same manner as physical handicaps, this fact does not seem to lessen the desirability of having available separate data on long-term Assistance cases, given the assumption regarding the maintenance of over-all full employment.

In addition to separate identification of National Assistance cases in the monthly statistics, the researcher would be assisted by the periodic provision of certain additional data regarding the applicant group as a whole, supplementing that now published quarterly in the *Ministry of Labour Gazette*. Such data would be of a qualitative nature designed to throw light on the characteristics of the unemployed and their attachment to the labour force.

Within the general body of applicants there are a certain proportion who, although in general able to qualify for Unemployment Benefit, have relatively great difficulty in being reabsorbed into regular employment. Such persons tend to fall into two not entirely mutually exclusive categories:

1. Persons unemployed for lengthy periods;
2. Persons subject to frequent spells of unemployment.

In many cases long-duration or intermittent unemployment will be due to lack of work opportunities in the area in which the claimant resides. In others however, personal characteristics, e.g. age, marital status etc., and occupational attachment may be important factors. Frequently the combination of a declining local industry with personal factors such as advancing age and/or family responsibilities, which inhibit movement to another community, will form the basis for the problem.

In studying long-term and recurring unemployment the researcher is aided by having available as much detailed information as possible regarding the personal characteristics and occupational attachment of the unemployed in relation to duration and frequency of spells of unemployment. Only with such qualitative information can labour force attachment and prospects for reabsorption into regular employment be assessed.

It is realised that the collection of statistics is not the primary function of Employment Exchange personnel and that the factor of

cost and space in official publications restricts the amount of qualitative data that it is practical to collect and publish. Even so it would be helpful if certain additional information could be published. At quarterly intervals at present information is published giving analyses of the wholly unemployed according to (*a*) occupation, and (*b*) duration of unemployment. Every other quarter the latter analysis is combined with a detailed analysis by age. This could be supplemented by data regarding such matters as duration of employment by occupational attachment and spells of unemployment experienced by the same individual during the year. It would also be useful to know the number and characteristics of persons who claim benefit regularly every year.

In using tables showing the characteristics of the unemployed the researcher would be assisted by having available in the same table comparable data for the employed as well as the unemployed. This procedure would reveal, for example, in data regarding age distribution, not only the importance of a particular group in the unemployed total, but also the significance of the group in relation to its representation in the total employee segment of the insured population.

Official statistics of short-time employment are available monthly for manufacturing industries (other than shipbuilding and ship-repairing), and show the number of employees working short-time or overtime and the average number of hours lost or worked overtime. Unemployment Benefit is payable to persons on short-time if unemployed at least two days within any period of six consecutive days excluding Sundays, but such persons are not included in the official count of the temporarily stopped unless one of the two days off happens to be the Monday on which the monthly count is taken.

It has been contended that short-time, or partial unemployment, represents a form of 'hidden unemployment' and that the official count of unemployment should be adjusted to take time lost from this source into consideration. Thus H. A. Turner has recently written that: 'Reduced hours now seem to be a more important consequence of bad trade than full unemployment and their effect is not fully recorded by available public statistics.'[12]

As A. F. Burns and Wesley Mitchell have demonstrated[13] for the manufacturing industry in the United States, a statistical series based on persons working less than full time is a useful indication of change in the economic climate. In some respects such a series might be regarded as a more sensitive indicator than statistics of those totally unemployed. Thus partial unemployment may forecast the direction of economic movement while total unemployment is the result of a decline in production which may have been going on for some time before large-scale layoffs occurred.

Given that changes in the amount of short-time are a valuable index of change in economic activity it might be argued that short-time

should not be expressed in terms of units of full-time unemployment. In the case of persons on short-time the need is for more work on existing jobs, while the totally unemployed require to be placed in employment. As the needs of the two groups differ, so do the policies required to meet these needs; and for policy action a variety of information is generally more helpful than a single set of figures.

Aside from the question of the desirability of converting time lost through short-time into units of full unemployment, such a conversion procedure cannot readily be done with precision using information available from current establishment surveys.[14] To effect the conversion of time lost through short-time into full-time unemployment a frequency distribution of persons by hours worked would be required and this would imply the use of a direct sample survey of the working population. It does not appear that a survey of this type is required for the overall measurement of unemployment in Britain and, bearing in mind the expense of such an undertaking, it is doubtful whether it could be justified on any other grounds.

Acknowledgements

Helpful discussions with officers of the Statistics Department of the Ministry of Labour and National Service and the Ministry of Pensions and National Insurance, and with Mr H. A. Turner of the University of Manchester, are gratefully acknowledged.

REFERENCES

[1] *Annual Report, Ministry of Pensions and National Insurance* (1957), Table 15, p. 57. Cmd. 493. HMSO, London.
[2] Turner, H. A., *The Times*, 19th November 1958.
[3] Turner, H. A. (1955). 'Measuring unemployment', *J. R. Statist. Soc.*, A, **118**, 39.
[4] Titmuss, R. M., *The Times*, 4th December 1958.
[5] Slichter, S. *et al.* (1950). 'How much unemployment', *Rev. Econ. and Statist.*, **15**, 76.
[6] Turner, H. A. (1959). 'Employment fluctuations, labour supply and bargaining power', *The Manchester School*, **27**, 175.
[7] United States Bureau of the Census (1948). *Labour Force Memo. No.* **3**.
[8] Reynolds, L. G. *et al.* (1948). *Survey of Contemporary Economics*, p. 271. Blakiston Co., Philadelphia.
[9] *Report of the National Assistance Board* (1956), App. 8. Cmd. 181. HMSO, London.
[10] *Idem* (1957), p. 19.
[11] *Monthly Digest of Statistics* (December 1956-1959), Table 29. HMSO, London.
[12] Turner, H. A., *The Times*, 2nd December, 1958.
[13] Burns, A. F. and Mitchell, W. (1938). *Statistical Indicators of Cyclical Revivals.* Bulletin No. 69, National Bureau of Economic Research.
[14] Estimates of this sort have, however, been prepared. See: Turner, H. A. (1955). 'Measuring unemployment', *J. R. Statist. Soc.*, A, **118**, 28.

MEETINGS OF SECTIONS AND LOCAL GROUPS OF THE ROYAL STATISTICAL SOCIETY

INDUSTRIAL APPLICATIONS SECTION

2nd October 1959. Inspection in the Paper Industry: I. F. Hendry
6th November 1959. Can Statistics Help the Manager? P. J. Stanley
18th December 1959. Half-day Conference: Experimentation in Theory and Practice: G. A. Barnard (Chairman), D. R. Cox, E. J. Whyte, and P. G. Moore

GENERAL APPLICATIONS SECTION

14th October 1959. The Study of Elections: H. C. Mackenzie
11th November 1959. Some Aspects of the Radcliffe Committee's Report: H. B. Rose (Joint meeting with the Institute of Actuaries Students' Society)
11th December 1959. Pitfalls in the Collection of Agricultural Statistics: H. Palca

MEDICAL SECTION

29th September 1959. Mortality Studies. Work in Progress: R. A. M. Case
27th October 1959. Some Problems in the Estimation of Familial Aggregation of Disease: W. Haenszel
24th November 1959. Recent Trends in Mortality: R. D. Clarke

BIRMINGHAM AND DISTRICT GROUP

7th October 1959. Accidents and Road Layout: R. F. Newby
4th November 1959. Closed Queues: F. Benson
2nd December 1959. Statistical Methods in Auditing: P. N. Wallis

BRISTOL GROUP

15th October 1959. Ways of Measuring Uncertainty: G. A. Barnard
19th November 1959. Paired Comparisons with Particular Reference to Social Investigations: L. T. Wilkins
17th December 1959. Delays on a Two-Lane Road: J. C. Tanner

EDINBURGH GROUP

1st December 1959. Statistical Methods in Brewing Research: I. A. Preece

GLASGOW GROUP

17th December 1959. Statistical Problems in Machine Minding: F. Benson

LEICESTER GROUP

1st October 1959. The Indolent Plotter: D. Goldberg
19th November 1959. The Marriage of Statistics and Computers in Operational Research: K. D. Tocher

MERSEYSIDE GROUP

10th November 1959. Selling Statistics to Management: R. H. S. Lesser
8th December 1959. Evolutionary Operation in Practice: G. A. Coutie

NORTH-EASTERN GROUP

14th October 1959. Fractional Factorial Experimental Designs: G. A. Coutie

Mr Coutie described an experiment in which nine factors had been varied in order to locate a source of unusual variability in a test-criterion. The design chosen had been a fractional factorial. He went on to discuss the merits and deficiencies of such designs and, in particular, described how factorial designs could be used, in conjunction with balanced incomplete-block designs, to generate a new class of second-order rotatable designs involving only three levels of each variable. This class of designs was compared with the equivalent set of composite designs based on full or fractional factorial experiments. Finally, the immense value of simple two- or three-factor two-level factorial designs in Evolutionary Operation was stressed.

11th November 1959. Practical Problems in Error-Correcting Codes: K. D. Tocher

9th December 1959. Problems in Air Pollution Measurement: D. Kerridge

SHEFFIELD GROUP

29th October 1959. Experiments in Telepathy: S. G. Soal

26th November 1959. Rehearsal for Business: J. Laski

SOUTH WALES GROUP

8th October 1959. Industrial Simulation by Computer: K. D. Tocher

12th November 1959. Normalising Transformations: M. H. Quenouille

8th December 1959. Experiments in Telepathy: S. G. Soal

BOOK REVIEWS

G. L. S. Shackle: *Economics for Pleasure.* Cambridge University Press, 1959. Pp. x+269. 21s.

Dip into this volume almost anywhere and there is, in fact, real pleasure to be obtained. Ideas of economic theory are explained in a fascinating way. There are no diagrams or mathematics; algebra is avoided so deliberately that we read '[one *divided* by (one *minus* the marginal propensity to consume)] and wonder whether the parentheses are not here mathematical symbols which should be avoided, as are 1, ÷ and −. (Figures are used elsewhere in the text and even the triple parallel lines which mean 'identically equal'.)

The headings of the eight 'books' and the five chapters in each book comprise a good syllabus for a course in economics. The books are on value, production, income, distribution (of income), employment, finance, government, and (international) trade. In small doses the chapters instantly appeal and cover their subjects well. Forty chapters at half-an-hour each give 'Economics for pleasure' if the half-hours are well spaced. Take them too close together, however, and there is danger of indigestion. The beautiful symmetry begins to pall and then to annoy. The subject becomes an abstract science, a process of logical deduction removed from reality. The volume, excellent in detail, is disappointing in the aggregate. It rouses in me the same feelings that I had when I changed from Workers' Educational Association classes in economics to the University approach, feelings that I thought had long ago passed completely. We learn how the shopping housewife balances an ounce of butter with three or four ounces of sugar at the margin; excellent logic. But we do not hear how she imposes her decision on the members of the household. Does sugar just run out at Friday breakfast or does she stop providing home-made sweet cakes? Where does the freedom of choice at given prices come in for the ordinary members of the household? Can one imagine the conversation whilst the decision is being made? Surely there are many things the normal household consumes to satiety? People who stop eating before they are satisfied do it for reasons of health or appearance, not because of expense. Or consider the careful balancing of extra leisure against the reward for extra work at the margin and compare with the way individuals accept working hours as given but often like overtime. Are we to believe the diminishing marginal utility of income? Why not submit an alternative proposition that the marginal receipt of income is often the most important? The remainder of income is absorbed in living expenses, it is the last element that provides life with its thrills and pleasures. Most people settle to a level of living determined by, shall we say, last year's income and this gives no great pleasure, it is just a sort of maintenance cost; the great pleasure is provided by any additional earnings or windfall. Arguments of this kind are not fully believed, but too much economic theory which sounds like thought divorced from reason does encourage search for plausible alternatives.

Further consideration of the subject leads one to conclude that the analysis in this volume is a necessary but a far from sufficient explanation of human behaviour when faced with the impossibility of achieving satiety in

desirable things. It does not explain fully the current use of the given collection of productive resources or why they are not always being fully employed. And these are the things that the volume tells us economics has to attempt. Some of the questions that seem very important and are not adequately dealt with are given below.

Why do many firms prefer to spend on advertising instead of reducing prices? What is it that makes tastes change to such an extent that they often swamp the price influence on demand? Advertising is not discussed, though there is mention of each firm having its own distinctive product, whether motor scooter or kind of chocolates, and this being a source of imperfect competition. Is it realistic to dismiss rationing in five lines whereby 'quantities available give less satisfaction than they might'? Finally, is government influence on the use of resources adequately dealt with in the chapter on planning when it is remembered that at present government expenditures and transfers out of taxation amount to as much as one-third of national product, and all such payments influence the current use of resources? It is not sufficient for the author to say he believes that the great aggregates should be planned and that everyone should have provided the essentials of decent life and the opportunity to develop his whole personality. In any case these comments seem like moral judgements rather than explanations of what happens.

This volume is intended to appeal to the general public as an explanation of the working of the economic system. As such it can hardly ring true to the layman and will only confirm the layman in the hatred, ridicule, and contempt mentioned by Professor Shackle in his opening sentence. This will be a pity, for the book can be read with great pleasure and profit by the student of economics who looks upon this sort of analysis as dealing with part, but only part, of the factors involved. H. S. BOOKER

International Journal of Abstracts—Statistical Theory and Method: W. R. Buckland (General Editor). Published for the International Statistical Institute by Oliver and Boyd Ltd., Edinburgh. Annual subscription £5 (or $16); single copies 30s. (or $4.50).

The aim of this journal of abstracts is to give complete coverage of papers in the field of statistical theory and new contributions to statistical method; papers which only report applications or examples of existing statistical theory and method will not be included. There are about two hundred and fifty journals which are wholly or partly devoted to the field of statistical theory and method and which form the raw material. Others may have to be added later. Some of the journals will be completed abstracted; others will be abstracted on a virtually complete basis; and the remaining journals, the vast majority, will be regularly scanned and relevant papers abstracted. In addition reports of research stations, conferences, symposia, etc., will be covered. The abstracts are up to about 400 words long, and so give scope for some detail and the inclusion of equations where necessary.

A classification into twelve main sections has been adopted; these are printed on paper of different colours to aid in identification. Each section is then further subdivided. Each abstract is then given two two-figure code numbers: the primary one is determined by its basic topic and the secondary one by its most important cross-reference. The journal is so arranged that it may be (*a*) left intact as a periodical; (*b*) filed in page-form according to the main sections of its classification; (*c*) split into pages and guillotined so that

each abstract may be kept separately, either pasted on cards or filed in binders (appropriate holes are already punched).

There are about 160 abstracts in this first issue and the eleven journals which are abstracted on a complete or virtually complete basis account for about three-fifths of the total. This is to be expected, but it is even more important to stress the fact that the remaining 40 per cent of the abstracts are from journals that the majority of statisticians would be unlikely to see regularly. This is quite a high proportion.

All in all, the journal seems likely to prove a very useful information service, and it should be particularly appreciated by the professional statistician.

D. G. BEECH

Norman T. J. Bailey: *Statistical Methods in Biology*. London: English Universities Press, 1959. Pp. ix + 200. 25s.

In recent years it has become clear that statistical methods are a very great assistance to biologists. In many branches of biology they are essential. Hence there is a need for books explaining elementary statistics to biologists in a simple straightforward manner, without that air of mystery which so often surrounds the subject in standard textbooks. Some books attempt this, but none so successfully as Dr Bailey's.

The book deals with all the standard concepts and methods: normal, Poisson and binomial distributions, *t*-tests, contingency tables, total and partial correlation and regression, the analysis of variance, experimental design, factorial experiments, randomisation, and sampling procedures. A very useful appendix gives a summary of statistical formulae, with a guide indicating for any ordinary type of data the appropriate scheme of calculation and the tables (either in this book or elsewhere) needed for the final interpretation. The explanations throughout the book are lucid, and the comments shrewd and sensible.

There are a few minor lapses from this high standard, which are, however, almost inevitable in a first edition. For example, the definition of probability on p. 10 is so worded that it fails to give any clear definition at all. The addition and multiplication laws of probability are used on pp. 11 and 12 without much explanation, and the symbol $n!$ is introduced without definition on p. 12. In addition, the author several times recommends the use of Bartlett's test for homogeneity of variance, but only mentions in a rather inconspicuous paragraph that it is sensitive to deviations from normality. The section on contingency tables might be improved by an account of Woolf's G test, as well as χ^2. The example in Table 2, p. 12, would be more plausible in animal than in human genetics (for human families of this kind lacking albinos would be undetected). The recommendation in the second paragraph of p. 129 seems to suggest that it is wrong to make estimates too accurate. Surely it depends on how these estimates are obtained, and what use is to be made of them? An accurate estimate may be useless in a significance test, if its estimated standard error is too big, for a possible significance may then be missed. But in an estimation problem this objection may be less important. These criticisms are, however, only concerned with minor details.

In short, this is just the book for the beginner. It can also be recommended to those who have already learnt to apply statistical methods somewhat mechanically, but wish to understand them better.

However, it is interesting to look at this book from another angle. Here

we have set before us with great clarity a panorama of modern statistical methods, as used in biology, medicine, physical science, social and mental science, and industry. How far does this show that these methods fulfil their aims of analysing the data reliably, and how many gaps are there still in our knowledge? There are four main types of problems which arise, namely those dealing with proportions, means, variances, and degrees of association. In the case of means and proportions the book shows that the position is generally very satisfactory. There are some controversial points, but these make little difference in practice: there is also need for further study of the effect of large differences in variance on some types of analysis of variance procedure. Bailey is very conscientious in giving warnings of the need for caution in interpretation in certain circumstances, such as non-normality in t-tests or small numbers in contingency tables. But it is unfortunate that at present these warnings must be rather vague: the reader will want to know how much allowance to make for failure in the assumptions. (Perhaps non-parametric tests might be useful?) Comparisons of variances are more difficult, since the F and Bartlett tests depend rather critically on the assumption of normality. In the case of measures of association, there is already a formidable mathematical apparatus of regression and correlation coefficients. In practical problems it is often less clear which method to use, and some discussion might be helpful. For example, is a problem most appropriately solved by using the regression of y on x, or that of x on y, or by some intermediate curve? The aim of the experimenter may be concerned here, as well as the mathematical properties of the statistics.

One feature which can puzzle an outsider, and which requires much more justification than is usually given, is the setting up of unplausible null hypotheses. For example, a statistician may set out a test to see whether two drugs have exactly the same effect, or whether a regression line is exactly straight. These hypotheses can scarcely be taken literally, but a statistician may say, quite reasonably, that he wishes to test whether there is an appreciable difference between the effects of the two drugs, or an appreciable curvature in the regression line. But this raises at once the question: how large is 'appreciable'? Or in other words, are we not really concerned with some kind of estimation, rather than significance?

Another feature which requires more explanation is that of using the result of a supplementary significance test to decide the type of analysis. For example, if the variances in two samples do not differ significantly, we use a t-test to compare their means: otherwise some modification is necessary. But what magic is there in some arbitrary significance level, in that the method of analysis should depend discontinuously on whether it has been exceeded or not? Sometimes this difficulty can be resolved fairly simply, as in the case, considered on p. 95, of a test for independence in a number of 2×2 contingency tables which may not be homogeneous. For each table we calculate the expectation and variance of the entry in the top left-hand cell, using the exact distribution. The observed numbers in this cell in all tables are then added together, and so also are the expectations and variances: the square root of the total variance is the standard error of the total. Finally the difference between the total observed and expected is compared with the standard error, to see if there is significant disagreement. Provided the numbers in the individual tables are not too small, this gives a satisfactory test. (It remains valid even if some tables have small numbers, but loses efficiency).

Here are at least some of the gaps which must be filled before the perfect book on statistics can be written. Meanwhile we must be content with a good approximation, and Bailey's book provides that.

CEDRIC A. B. SMITH

James H. Batchelor: *Operations Research: An Annotated Bibliography*. Saint Louis University Press, 1959. Pp. x + 866. $10.

It would seem that the two desiderata of any bibliography are an adequate degree of selection of material for inclusion and a comprehensive system of cross-referencing. In this case, the unnecessary inclusion of irrelevant articles has been compensated to some extent by a very useful index.

The criterion for inclusion in this book has been 'that the author or editor has indicated that the work concerns operations research, or that it has been cited in support of such work'. This has resulted in some curious references such as:

1405 R. S. Gander—Appreciation
JORSA, Jun 1956, 4:3:361.
Appreciation for the generous response to requests for papers.

On the other hand, it has entailed quite a number of omissions. Thus, for example, there is no mention of the work done by Lloyd S. Shapley for the RAND Corporation on the Theory of Games.

So that the reader can judge for himself, I have taken the first ten four-digit random numbers from Fisher and Yates's Statistical Tables that are less than or equal to 4195, the number of references, and have given below the titles, without annotations, of the corresponding references:

0347 Edward G. Bennion—Capital budgeting and game theory.
3661 H. Steinhaus—O prognozie (On prediction).
2616 Carson B. McGuire—Steam locomotive availability and terminal facilities.
1410 H. Newton Garber—Information handling at Westinghouse's Transformer Division.
1457 Weldon B. Gibson—Techno-Economics—a revised check and balance for management of industrial research.
2042 H. G. Jones, A. M. Lee, D. T. Steer—Oxygen for steelworks use.
3732 Leonard W. Swanson—Operations research in business and industry.
2707 Modern Industry—Putting your problems through college.
3606 Vernon L. Smith—Economic equipment policies; an evaluation.
1676 B. Harris, A. Hauptschein, L. S. Schwartz—Minimum cost decision-feedback systems for detecting signals perturbed by additive gaussian noise.

The annotations are clear and concise; the printing (mimeographed) is, however, a little difficult to read. The references are arranged in alphabetical order of authors' names and an index is provided in a single alphabetical order with entries by author, title, and subject. The references cover the period up to the end of 1957.

It should be noted that there exists another bibliography on operations research which is published by the Operations Research Society of America, covering a rather more relevant area of work; it does not, however, contain summaries of the contents of the articles. Although the industrial statistician will probably be content to thumb through back-numbers of the relevant journals, there is little doubt that either one of these bibliographies could ease the burden of many librarians.

A. MUIR

John E. Freund and Frank J. Williams: *Modern Business Statistics*. London: Pitman, 1959. Pp. xv+539. 50s.

This should be a good textbook for a beginner, particularly for one whose mathematics is weak, rusty or non-existent. Each chapter is provided with several sets of exercises and with lists of books for reference or further study. With this book the student should fairly quickly obtain some knowledge of the whole range of statistical methods and their inter-relationships.

The chapters on distributions, probability, and sampling follow a conventional but well-knit pattern of exposition; a relatively advanced level is achieved by very simple methods of treatment. Two interesting chapters on tests of hypotheses provide the student with some useful tools and give a lead in the study of the analysis of variance. Linear regression and correlation are then introduced; the coefficient of correlation is dealt with at some length and rank correlation, multiple and partial correlation, and contingency are dealt with briefly.

A fair amount of space is given to time series analysis and to index numbers. The chapters on index numbers are very readable and apart from dealing with the usual theoretical problems, they contain an unusual amount of practical wisdom on the subject. The authors, however, confine their discussion mainly to simple price indices; they thus lose the opportunity of demonstrating the relationship between price and volume indices of the Laspeyres and Paasche types and of bringing to light one of the chief problems in indices of production—that of measuring the *net* output of industries. In these days even an elementary textbook should introduce the student to the concepts of 'net' and 'gross', which are important not only in connection with indices of production and the weighting of certain wholesale price indices, but also in certain branches of national income statistics, a subject in which businessmen are taking an increasing interest.

The appendices are rather a mixed bag: I Pictorial Presentations; II Quality Control; III Calculations with Rounded Numbers; and IV The Use of Logarithm and Square Root Tables. Perhaps the collection would not have seemed so odd if the material on quality control had been removed to the body of the book. Condensed versions of the usual statistical tables follow and, in a final section, answers are provided for about two hundred of the exercises.

Throughout the book the text, formulae, tables, diagrams, and charts are presented in an agreeably clear type. T. M. RIDLEY

Harold F. Dodge and Harry G. Romig: *Sampling Inspection Tables, Single and Double Sampling* (Second edition, revised and enlarged). New York, John Wiley and Sons; London, Chapman and Hall, 1959. Pp. xi+224. 64s.

The original edition of these Tables in book form appeared in 1944 in response to considerable demand for reprints of the inspection tables published in the *Bell System Technical Journal* in 1941. Since then the 'Dodge-Romig Tables' have become a household word among inspection engineers and others. The tables were pioneers, not only in the sense that they publicised new concepts and stressed the idea of rectification, but also in applying a consistent system to a wide range of manufacturing conditions. In later years other inspection tables, employing other systems based on different assumptions, have been published and there is no reason to believe that the process will cease. Hence the present edition faces more 'competition'. Each system has its advocates and it is not my intention to compare the

systems here. Suffice it to say that Dodge and Romig provide a consistent system whose primary object is to satisfy the quality requirements with the minimum inspection.

Those familiar with the somewhat austere appearance of the wartime edition will be surprised at the present opulent volume; it is much larger and the layout is greatly improved. The increased size is partly due to the improved layout and partly to the inclusion of additional material. The major portion of this consists of three appendices giving operating characteristic curves and a chapter explaining them. The charts giving these curves occupy over one hundred pages, with four charts per page and several curves per chart, and are very handsomely reproduced. They provide additional information for the users of the sampling plans, and extend the usefulness of the volume to those who wish to apply non-rectifying inspection and to have some latitude in their choice of a sampling scheme. In view of the lavish scale of the book, the price is not unduly high.

D. G. BEECH

PUBLICATION RECEIVED

A. R. Ilersic: *Statistics* (12th Edition). London: H.F.L. (Publishers) Ltd., 1959. Pp. xiv + 467. 30s.

The first eleven editions of this book were published under the title: *Statistics and Their Application to Commerce*. A review of the eleventh edition appeared in our November 1957 issue. The book has now been completely revised; for those teachers and students who may have used earlier editions a brief note of the changes in this edition may be helpful.

The chapters on social statistics and vital or demographic statistics are new, and those on sampling theory and significance tests have been largely re-written. There has been some re-arrangement of chapters and some have been substantially revised. The material on economic and business statistics has been brought up to date.

The book is designed primarily for students taking external University or professional examinations in which statistics is a subsidiary subject.

THE ECONOMIC INCENTIVE PROVIDED BY SAMPLING INSPECTION

I. D. HILL
War Office

In this article Mr Hill considers two aspects of sampling inspection—the curative and the incentive. He discusses the importance of the incentive aspect and indicates a possible line of approach to a theory of sampling inspection which takes this aspect into account.

Introduction

The theory of sampling inspection, as usually developed, seems to be devoted almost entirely to the question of what I call the curative aspect—that is to say, it assumes that the quality of production that may be offered for inspection is independent of the sampling scheme being used, and that the job of the sampling scheme is to 'cure' bad quality production, when it arises, by rejecting the bad batches so that they do not reach the consumer. An alternative view is that one of the principal jobs of a sampling scheme is to prevent rather than to cure, by influencing the quality that may be offered. This incentive aspect has received very little attention, in print at any rate.

The concept was used in a paper by Whittle[1] in 1954. He illustrated the main idea as follows: 'When a tramway inspector boards a tram at random and examines the passengers' tickets, he is not trying to estimate the number of unpaid fares. His true function is not even to detect unpaid fares—the amount of money he recovers directly would certainly not pay his salary. The real purpose of the inspection is to inhibit transgression. The mere knowledge that "sampling" is being performed deters people from carelessness or worse—the more intense the sampling, the greater the deterrent effect.'

Whittle considered only the case where the inspection is of individual items rather than of batches of items, and there is a fixed fraction of items that will be inspected. He was then able to make a plausible assumption about the way in which the incoming quality might vary according to the sampling fraction. I assume, although he did not actually say this, that the punishment for a defective item is simply the rejection of that item. If this is so, then the deterrence must increase as the sampling fraction increases and the best incentive, as well as the best curative effect, will be given by 100% inspection.

But if, as is usual in industrial sampling inspection, the items are presented in batches, then 100% inspection and rejection of individuals may still give the best curative effect; but sampling inspection with rejection of batches if an acceptance number is exceeded will, in general, give better results from an incentive point of view.

This is seen again and again in practice. The quality being offered for 100% inspection may continue at a fairly constant level almost indefinitely, but as soon as 100% inspection is replaced by an appropriate sampling scheme the quality being offered improves. We thus get what appears at first sight to be a paradoxical situation—that the inspector can achieve a better product for the consumer by doing less work. If this is so, and I am sure that it is, then it is clearly of great importance in industrial inspection, and yet so much of the discussion on acceptance sampling concentrates entirely on the curative aspect to the exclusion of the incentive aspect.

One reason for this is that the curative aspect is an essentially simpler thing about which to make up a theory. In particular it is not dependent upon how, where, and when inspection is being performed; a given sample size and acceptance number have the same effect, curatively, whether inspection is of batches immediately after production to decide whether to buy the goods from the manufacturer or not, or of batches which have been bought in the past (and have been in store for years) to decide whether to keep them or scrap them. From the incentive viewpoint, however, these are very different situations. In the first the decision on any particular batch has the chance of influencing the quality of the following batches; in the second there is no such chance, and incentive ideas cannot apply at all—only curative ones are relevant.

Another fact which makes the theory of the curative aspect simpler than that of the incentive aspect is that in the former we deal solely with objective probabilities and can say, with certainty, what is bound to happen in the long run. In the latter we have to deal also with subjective probabilities and can say only what is likely to happen.

The Case for Using the Incentive Aspect

The case for using the incentive aspect, if it can be done, becomes clear when we consider how the quality received by the customer is influenced by the process curve (or prior distribution) and by the sampling plan acting according to its operating characteristic curve. It is quite a common fallacy to assume that the eventual quality depends mainly upon the sorting ability of the sampling plan, and that the consumer can discover what quality he can expect to receive by studying the OC-curve. In fact this is far from the truth: if the manufacturer is making a fairly constant quality product the consumer will receive very much that same quality. If the manufacturer's quality varies a good deal the sampling can improve the average quality received, but even in this case the distribution as received by the consumer will nearly always bear a distinct resemblance to the prior distribution.

To consider specific examples I shall first take the case where the incoming quality is absolutely constant: i.e. the prior distribution consists of 100% frequency concentrated at a single point. This is not

of course, a very realistic example but it is an instructive one. (The only way, I imagine, in which it could happen would be if a manufacturer had, say, twenty machines all making the same product, and nineteen of these turned out 100% good product, the other one 100% bad product. If this situation could continue unnoticed, and if the batches were formed by mixing equal proportions of the product from the twenty machines, then every batch would be exactly 5% defective with no variability.) In this situation the outgoing quality will be determined solely by the incoming quality—whatever the incoming quality may be, the outgoing quality will be exactly the same, no matter what sampling plan is adopted. The number of batches per unit time will be less, since some will be rejected, but for the moment we are concerned solely with the *distribution* of quality, not with the numbers, and this will be unchanged.

I am assuming here that when a batch is accepted the sample is accepted as well as the remainder of the batch. This is usual practice where the inspection is non-destructive. It is however also usual (and commendable) practice to reject any defective individuals found in the sample even though the batch as a whole is accepted. In general this makes only a negligible difference and I am ignoring it. I am also ignoring the possibility of rectification of rejected batches.

If the inspection is destructive then, under conditions of absolutely constant quality, a sampling plan of the usual type will actually worsen the quality received by the consumer, since a smaller number of defectives in the sample indicates a greater number left in the remainder of the batch. From a curative point of view it would be better in these circumstances to reverse the usual type of acceptance rule, and accept only if at least some given number of defectives is found in the sample. It can be imagined, however, what the effect of such a rule would be from the incentive viewpoint; is it believable that the quality would continue constant for long? In practice, of course, completely constant quality does not happen so this point is only of academic interest.

The nearest, I think, that we could ever expect to get to constant quality would be the situation where there was a constant probability that each individual article would be a defective. This situation leads to a binomial distribution of quality from batch to batch. If the inspection is destructive, this situation is the equivalent of completely constant quality from batch to batch for non-destructive inspection. The accepted quality will have precisely the same distribution as the submitted quality whatever sampling plan is used. This is the result known as 'Mood's Theorem'.[2]

With a binomial prior distribution and non-destructive inspection Mood's theorem still holds approximately when the sample is only a small proportion of the batch. The sampling plan does make some change in the distribution, but for (sample size)/(batch size) ratios of the order generally found in practice the difference is only very slight.

As a numerical example I have made some calculations for the case:

Batch Size = 1000, Sample Size = 100, Acceptance Number = 2. Figs. 1 and 2 show the situations when the submitted quality is respectively 1% and 5% defective with a binomial distribution. (I have used the Poisson distribution for the submitted quality instead of the binomial. This is more convenient and is quite accurate enough.) There is a difference between the submitted and accepted distributions, but in each case the difference is marginal. The average values change from 1% defective submitted to 0·98% defective accepted, and from 5% defective submitted to 4·67% defective accepted.

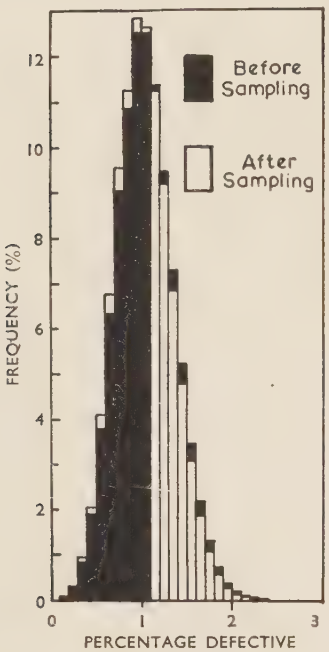


FIG. 1. Binomial frequency distribution for Batch Size 1000 and Process Average 1% defective, and the distribution of outgoing quality after applying the sampling plan: Sample Size 100, Acceptance Number 2.

Looking at either Fig. 1 or Fig. 2 we see how little difference is made by the sampling plan; comparing Figs. 1 and 2 we see what a large difference can be made by improving the production. Where the customer requires good quality it is clear that getting the production right is the important thing, and sampling inspection may be thought of as an aid to this rather than as a curative process to 'inspect good quality into the product'.

Now I must admit that this example cheats to some extent since it assumes binomial frequency distributions, which are the closest to absolutely constant production that

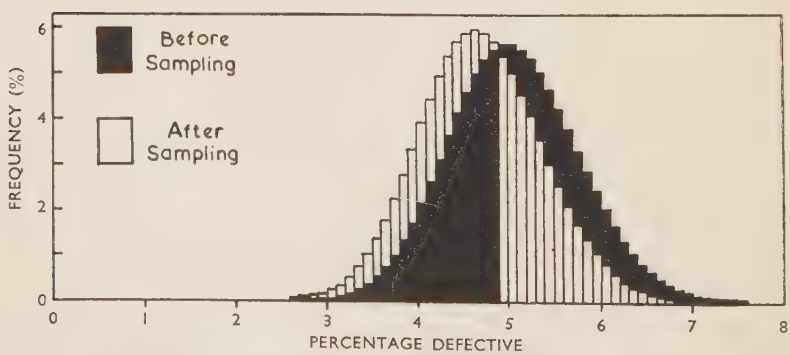


FIG. 2. Binomial frequency distribution for Batch Size 1000 and Process Average 5% defective, and the distribution of outgoing quality after applying the sampling plan: Sample Size 100, Acceptance Number 2.

we could hope to achieve. In practice process curves nearly always have a greater scatter than a binomial with the same process average, but as soon as we leave the binomial there is some doubt about what shape of distribution we ought to assume. Of the many suggestions that have been made the 'mixed binomial' model seems to have advantages. Any distribution can be represented as a mixed binomial by taking enough components, just as any planetary motion can be represented as closely as we please by taking more and more cycles and epicycles, but in either case the whole thing becomes rather meaningless if too many components are taken and, at any rate as a start, it is better for the sake of simplicity to restrict ourselves to the 'double binomial'.

A double binomial distribution has four parameters, N, p_1, p_2 and f, and is the resultant where a fraction f of the distribution of batches forms a binomial with parameters N and p_1 and a fraction $(1-f)$ a binomial with parameters N and p_2.

Fig. 3 shows the double binomial with $N = 1000$
$p_1 = 1\%$
$p_2 = 5\%$
$f = \frac{1}{2}$

Application of the previous sampling plan (Sample Size = 100, Acceptance Number = 2) to this distribution gives the distribution of accepted quality, also shown in Fig. 3. It is clear in this case, where the variability is considerably more than binomial, that the sampling plan is

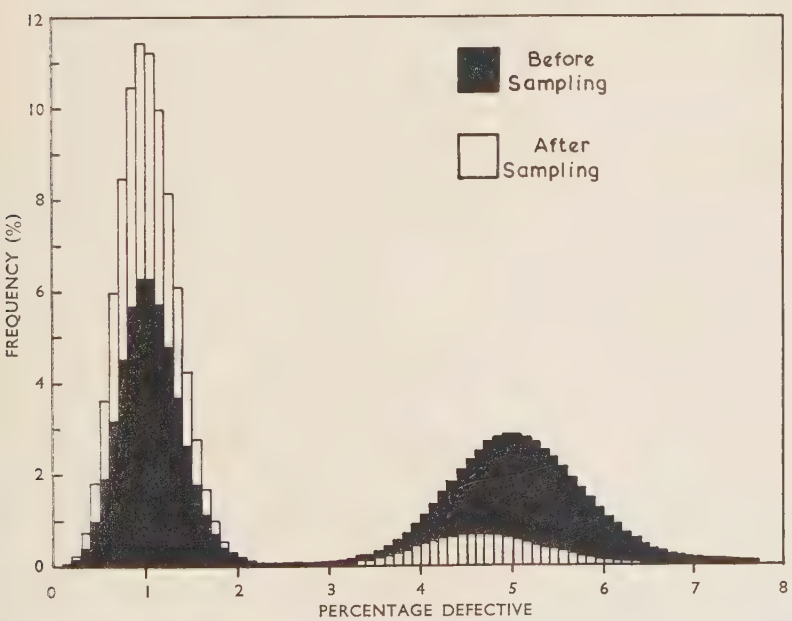


FIG. 3. Double binomial frequency distribution for Batch Size 1000 and equal components with averages 1% and 5% defective, and the distribution of outgoing quality after applying the sampling plan: Sample Size 100, Acceptance Number 2.

doing quite a lot of good, but again (comparing Fig. 3 with Fig. 1) how very much better results could be attained if, instead of tolerating the mixed binomial unchanged, the production could be tightened up to eliminate, or at least reduce the frequency of, the 5% component.

We have seen then that, at least in many cases, to aim at changing the prior distribution can be very much more effective than merely trying to sort out an unsatisfactory prior distribution. The question is 'Can it be done?' The answer is that it undoubtedly can and frequently is. Many of the sampling schemes in use make sense only if the incentive function is hard at work, but unfortunately, all too often, the very people who are using sampling in this way talk about it as if they were using it only curatively, as a sorting mechanism. This has the danger that the same schemes may get carried over, without realising the implications, to conditions where the incentive aspect is inoperative, with disappointing results.

The Requirements of an Incentive Scheme

I start with the following axioms:

(1) The prior distribution is entirely dependent upon the actions of the manufacturer;

(2) The making of goods is only a secondary aim of the manufacturer —a means to an end. His primary aim is the making of profits.

If these are accepted, then it follows from axiom (1) that if we are to have a good process curve, rather than trying to sort out a poor one, the manufacturer must take action. It follows from axiom (2) that he will not do so unless it is economically worth his while. It then becomes one of the principal tasks of inspection to see that the manufacturer has an economic incentive towards a satisfactory prior distribution. We need acceptance schemes which, if the prior distribution is not right, will cause economic loss for the manufacturer and at the same time will act as far as possible in a curative manner to improve the distribution, but if the prior distribution is right will reward the manufacturer with the profits he clearly deserves. If the prior distribution is right there is no necessity for the scheme to act in a curative manner as well, although some safeguard is evidently necessary as a precaution in case the distribution changes.

It is fundamental to the incentive function that it should be to the manufacturer's economic advantage to offer good quality and to his economic disadvantage to offer poor quality, which is achieved by having a scheme which accepts a high proportion of batches when quality is good, but rejects a higher and higher proportion if quality worsens. This might not seem to be anything worth saying, since operating characteristics always have this feature, but although the operating characteristic of a fixed sampling plan always does act in this way, it is common practice not to have a completely fixed sampling plan but to change the plan from time to time according to the sampling results. If we are intending the sampling to fulfil an incentive function

any such changes should be designed to reinforce the economic rewards and penalties, not to slacken them or to work in the opposite direction.

Judging by these requirements it is possible to examine some of the well-known tables of sampling schemes to see how they will act from an incentive point of view. Two of the best known tables are: (i) Dodge and Romig;[3] (ii) Military Standard 105A.[4]

It is not a criticism of Dodge and Romig's tables for the purpose for which they were designed to say that they seem to have some adverse features when considered for this different purpose. Considering the 'lot tolerance' tables, there will be a strong deterrent against quality as bad as, or worse than, the LTPD; but a good incentive scheme should do more than this. If there are two manufacturers both producing to the same specification with the same batch size, but with differing qualities, then an incentive scheme should always accept a higher proportion of the product from the better manufacturer than from the worse manufacturer. Dodge and Romig's tables do not do this, since they change the plan in use according to the quality offered so as to minimise the inspection costs. Fig. 4 shows the result of this policy in a typical case, namely,

LTPD	= 4·0%
Consumer's Risk	= 0·10
Batch Size	= 500

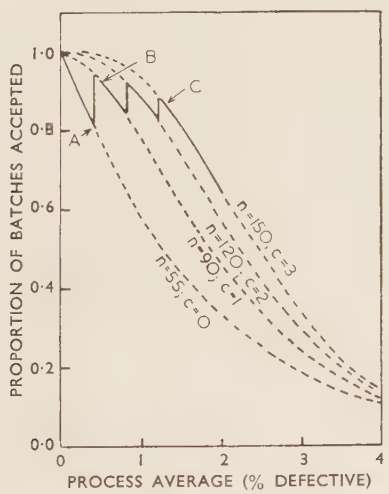


FIG. 4. OC-curves for Dodge and Romig sampling plans for Batch Size 500, single sampling, Lot Tolerance 4%, and binomially distributed incoming quality.

The OC-curves are shown for the four different plans given, each being used for a particular range of process average. This gives the resultant curve a zig-zag appearance; every time the curve tends to fall, there is a change of plan to bring it up again. If there are two manufacturers A and B producing binomial prior distributions with process averages of 0·4% and 0·5% respectively, A will have only 80·2% of his product accepted whereas B, giving slightly worse quality, will have 92·5%

accepted. This cannot be called an incentive to good production. Even a third manufacturer, C, who produces much worse quality, a binomial distribution with a process average of 1·3%, will have 86·7% of his batches accepted, which is considerably more than A, although his defective rate is more than three times as high. (I am assuming here that the binomial frequency distribution remains constant in each case and that the estimate of the process average is never so far away from the true value as to cause a change to a different sampling plan.) Dodge and Romig's AOQL tables give a similar effect, although the zig-zagging is much less severe.

Military Standard 105A also uses a system of changing from one plan to another when the process average changes, but the system of doing so is quite different, and here the changing does work so as to reinforce the natural action of the OC-curve rather than to slacken it. An example is shown in Fig. 5, which gives the OC-curves for an AQL of 4·0%, single sampling, and code letter I.

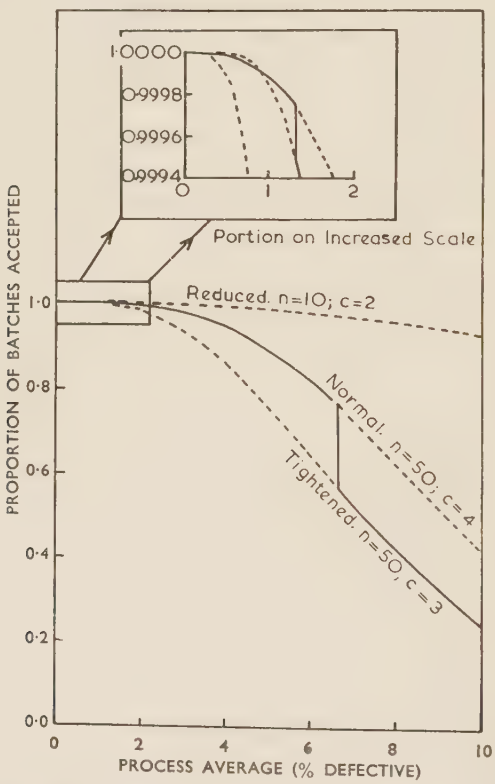


Fig. 5. OC-curves for Military Standard 105A sampling plans for AQL 4%, code letter I, single sampling, and binomially distributed incoming quality.

In making the calculations for Fig. 5, I have assumed that the process average is being judged on a 10-batch basis and that the estimated process average always comes out as equal to the true process average. This second assumption is very artificial, but the figure does

how how the change of plan works to steepen the OC-curve, instead f to level it off as Dodge and Romig's schemes tend to do.

There is one feature of Military Standard 105A which is anomalous: few of the Reduced Inspection plans give a smaller probability of cceptance at the AQL than the equivalent Normal Inspection plans. This might penalise a manufacturer for good work, and is not true to he spirit of the economic incentive. I said earlier that it was not a riticism of Dodge and Romig that their tables did not satisfy criteria hat they were not intended to satisfy, but I cannot here say quite the ame for Military Standard 105A. It has never, so far as I am aware, een stated that Military Standard 105A is intended as an incentive ype of sampling scheme, but I think it is clear, from the tables themelves and the way that they operate, that it is so intended. This lapse s therefore regrettable.

The Economic Approach

Over the last few years a number of people have contributed a fair mount of work on economic designs of acceptance sampling schemes n which the total costs of inspection plus wrong decisions are minimised. These have, so far as I know, always been from the curative ngle and minimise the costs for a given prior distribution, but, as Barnard[5] has shown, it is necessary to know the prior distribution in onsiderable detail to get a firm answer from such considerations.

The difficulties with economic schemes which assume a detailed nowledge of the prior distribution are: (i) that generally we do not ave this knowledge; (ii) that even if we do have the knowledge before he sampling commences the theory assumes that the distribution will ontinue unchanged; this is something it is very unlikely to do since the ampling will almost inevitably have an incentive effect, even though t was not designed with this in mind, and this effect is likely to lead to change in the prior distribution; (iii) that when such a change does ccur, if the sampling is left unchanged it will no longer be optimum, whereas if the sampling is changed as well the change will often be pposed to the incentive requirements, as happens with the Dodge and Romig schemes. Horsnell's sampling tables[6] are another example of his. Again I must say that I am not criticising these tables for not ulfilling a criterion they were not designed to fulfil; of course they would have a wider applicability if they could manage to meet all ossible criteria, but this is an impossible aim.

The solution of some of these difficulties might be to produce an conomic theory of sampling inspection which takes the incentive spect into account, and in the remainder of this article I indicate the nes upon which I think such a theory might be developed.

The Theory

The fundamental distinction from the standard economic sampling heories is that we have to make an attempt at judging what sorts of

prior distributions are possible, and what the relationship is between prior distribution and cost of production. For this preliminary attempt at a theory I have felt myself compelled to try something simple, even at the expense of reality, and I have made the following assumptions

(1) The prior distribution is a simple binomial.

(2) The cost of manufacture per article is $K/\sqrt{p}$, where K is a constant.

I must admit that this is simply a guess. I have no information on how costs do vary with quality and I have chosen this function merely because it is reasonably simple and appears to follow the sort of pattern that I should expect costs to follow. The pattern is shown in Fig. 6.

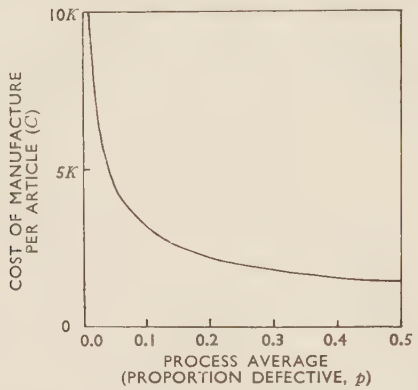


FIG. 6. The assumed relationship, $C=K/\sqrt{p}$, between process average and cost of manufacture for a binomial prior distribution.

and it will be seen that the curve is flat at the right-hand end but rises steeply at the left-hand end. This means that there is little difference in cost between producing at an average quality of 50%, 40% or 30% defective, but once better than 10%, improvements become more costly, 1% defective costing about 3 times as much as 10% defective, and the cost tending to infinity as the quality tends to perfection.

If costs do follow some pattern such as this it is obviously very important that a producer should be told what quality level is expected before he can quote a price for the job, and it is no good telling him that only perfection is good enough, since the price of perfection is infinite.

Let the cost of manufacture of an article be C

Then the cost of manufacture of a batch of size N is CN

Let the price paid for an accepted batch be LN

the average profit per batch be T

the sample size be n

the acceptance number be a

the proportion of batches accepted be P

$$\text{i.e.} \qquad P = \sum_{r=0}^{a} \binom{n}{r} p^r (1-p)^{n-r}$$

Then, assuming that a rejected batch is simply scrapped, and is a total loss to the manufacturer,

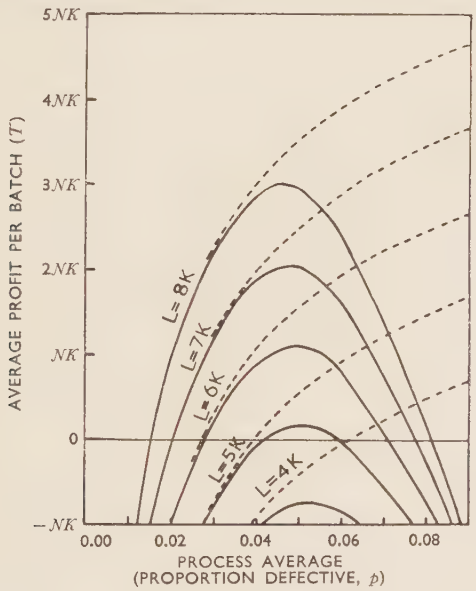
$$T = N(LP - C)$$
$$= N[LP - (K/\sqrt{p})]$$

Hence for a given N and K, T is proportional to $\frac{LP}{K} - \frac{1}{\sqrt{p}}$

In any given situation L and K will both be fixed but, for present urposes, it is only their ratio which is important. The manufacturer ill aim to adjust p so as to maximise T, and it can be shown (see ppendix) that, at the maximum, the value of p must satisfy the quation

$$\frac{1}{2\sqrt{p^3}} = \left(\frac{L}{K}\right)\frac{n!}{a!(n-a-1)!}p^a(1-p)^{n-a-1}$$

Fig. 7 shows (broken lines) the profit curves which would arise if ıspection were not performed and all product were automatically ccepted, i.e. if $P=1$ whatever p. These curves show that, on the ssumptions made, whatever the ratio between L and K might be, the



g. 7. Profit curves (shown as broken lines) where all product is accepted without inspection, given by $T=N[L-(K/\sqrt{p})]$ and (shown as continuous lines) where the sampling plan with Sample Size 150 and Acceptance Number 11 is applied, given by $T=N[LP-(K/\sqrt{p})]$.

anufacturer could always increase his profit by making worse quality. however a sampling plan is installed, $n=150$ and $a=11$ being taken an example, this is no longer true. There is a value of p at which the rofit is a maximum, and it does not pay the manufacturer to offer uality worse than this because of the effect of batch rejections. The rofit curves, where sampling is performed (continuous lines), show at the higher the price/cost ratio the greater is the profit attainable, nd the wider is the range of values of p for which it is possible to make profit. They also show that the value of p which maximises the profit

is relatively insensitive to changes in the price/cost ratio, and is alway round about 5% defective in this example. If the manufacturer doe try to adjust p so as to maximise his profits, it can be expected that th quality offered will be about 5% defective, in which case the samplin plan will accept between 90% and 95% of the submitted batches.

A fully developed theory of the incentive aspect of sampling woul need to be based upon genuine information of the relative costs o producing various prior distributions. Nevertheless, it is of interest tha the intuitive approach adopted here produces a situation in which th quality likely to be offered is quite close to 4% defective, since thi particular sampling plan is one of those given in Military Standar 105A for an AQL of 4%. Furthermore this is given as a 'norma inspection' plan; I have taken no account of the 'tightened inspection device. This device clearly improves the incentive value of the plan and would take the quality to maximise profit closer to 4% defective Military Standard 105A makes much more sense when examined i this sort of way, than when examined only with ideas of the curativ aspect in mind.

Acknowledgements

The original version of this paper was read at the Loughborough conference of the Royal Statistical Society in September 1959, and I should like to thank all those who took part in the discussion on tha occasion. They will recognise that some changes have been made a a result of their comments, and also that some have not been made i spite of their comments. I am equally grateful for the comments o both classes.

I also wish to thank Mr E. C. Fieller for many helpful comments.

APPENDIX

Derivation of the Equation which p must satisfy to maximise T

We have $T = N(LP - C)$

where $$P = \sum_{r=0}^{a} \binom{n}{r} p^r (1-p)^{n-r}$$

$$= \frac{n!}{a!(n-a-1)!} \int_0^{1-p} (1-x)^a x^{n-a-1} \mathrm{d}x$$

by the well-known relationship between the binomial distribution and the incomplete beta-function ratio.

$$\therefore \frac{\mathrm{d}T}{\mathrm{d}p} = N\left(-L\frac{n!}{a!(n-a-1)!} p^a (1-p)^{n-a-1} - \frac{\mathrm{d}C}{\mathrm{d}p} \right)$$

Hence, when T is a maximum,

$$L\frac{n!}{a!(n-a-1)!} p^a (1-p)^{n-a-1} + \frac{\mathrm{d}C}{\mathrm{d}p} = 0$$

In the particular case discussed,

$$C = \frac{K}{\sqrt{p}}$$

$$\therefore \frac{dC}{dp} = -\frac{K}{2\sqrt{p^3}}$$

Hence, at the maximum value of T,

$$\frac{1}{2\sqrt{p^3}} = \left(\frac{L}{K}\right)\frac{n!}{a!(n-a-1)!}\,p^a(1-p)^{n-a-1}$$

REFERENCES

[1] Whittle, P. (1954). 'Optimum preventative sampling', *J. Operat. Res. Soc. Amer.*, **2**, 197.

[2] Mood, A. M. (1943). 'On the dependence of sampling inspection plans upon population distributions'. *Ann. Math. Statist.*, **14**, 415.

[3] Dodge, H. F. and Romig, H. G. (1959). *Sampling Inspection Tables* (second edition). Wiley, New York.

[4] Military Standard 105A (1950). *Sampling Procedures and Tables for Inspection by Attributes*. Government Printing Office, Washington, D.C.

[5] Barnard, G. A. (1954). 'Sampling inspection and statistical decisions', *J. R. Statist. Soc.*, B, **16**, 151.

[6] Horsnell, G. (1957). 'Economical acceptance sampling schemes', *J. R. Statist. Soc.*, A, **120**, 148.

SOME PROBLEMS IN APPLYING STATISTICS TO GEOLOGY*

W. C. KRUMBEIN
Northwestern University, Evanston, Illinois

In this article Professor Krumbein points out that applications of statistics in some branches of geology are straightforward, but that geologists in general are faced with restrictions on probability sampling, with numerous variables that have to be 'sorted out' in terms of their relative importance, and with small-scale sources of variability that sometimes obscure broader trends in the observational data. Some ways in which geologists approach these problems are discussed to acquaint statisticians with an actively growing field for further application of statistical methods.

Introduction

One commonly thinks of geology as a qualitative science that uses numerical data sparingly if at all. With respect to the qualitative core of geological reasoning this opinion may be roughly correct, although the past half-century has seen an enormous increase of numerical data in geology. The numbers arise from quantification of earlier qualitative concepts, from development of new techniques in chemical and physical analysis of rocks and rock components, from controlled laboratory experimentation on geological processes, and from the almost explosive expansion in subsurface exploration for oil and gas that occurred within the past four decades.

Before the end of the last century geologists had begun to measure the sizes and shapes of grains in sedimentary deposits, and to estimate the mineral composition of rocks by identifying and counting crystals or grains under the microscope. Various statistical devices were used to summarise the data, and from this rather small beginning the use of statistics in some branches of geology expanded rapidly, especially in the 1920's and 1930's. (In this connection, however, see Fisher[1] on the use of a statistical argument by Lyell in the 1830's.)

Today statistical analysis is being applied in virtually all branches of geology, and the time lag between the development of new theory and techniques by statisticians, and their application to geological problems, is considerably shorter than it was in the past. The increasing availability of high-speed computers has also enabled geologists to use advanced methods of analysis that would be virtually impossible in the time required by ordinary desk-calculator methods.

Although many applications of statistics in geology are straight-

* This paper is a generalisation of some material presented in lectures delivered at several British Universities under the auspices of a Fulbright Senior Lecturer Award covering the period January-June, 1960. During this time the author's headquarters were at Queen Mary College of the University of London.

forward, there remain several broad aspects of the general problem of applying statistics to geology that require detailed consideration. These problems revolve in part around at least three considerations: severe sampling restrictions in some geological studies, the multiplicity of variables in even the seemingly simplest geological situation, and the high 'noise level' of some geological data.

The Nature of Geological Data

Elsewhere[2] I have reviewed some attributes of geological data, and part of this is repeated here to establish the framework of discussion. By far the greatest number of geological measurements are made on rocks and fossils formed by numerous interacting and simultaneously varying factors in geological environments that operated entirely without man's control, or even his presence. Geological factors operate at more than one level temporally and geographically, so that many small-scale fluctuations are superimposed on broader large-scale effects, giving rise to high degrees of variability in some geological measurements.

In his day-to-day work the geologist studies the end-products of natural processes that operated long ago. Moreover, he is in general limited in his observations to those places where rocks are exposed, as in outcrops, quarries, and boreholes. Even when a whole phenomenon can be examined, such as a present-day beach, it is physically impossible to measure every sand-grain or pebble. In practice the geologist examines the beach as a whole, and then selects one or more samples for more detailed measurement.

These attributes of geological observations—presence of numerous interacting factors, simultaneous variation among all or most of the factors, and limitation in large part to samples—are common to many if not most geological studies. It attests to the success of the science as a whole that from such data, evaluated mainly by qualitative methods, have emerged a number of well-established geological generalisations. More important in the present context, perhaps, is the fact that these attributes of geological data also demonstrate the essentially statistical nature of many geological observations. Statistical reasoning, whether formal or informal, is implicit in some aspects of virtually all geological studies.

The Sampling Problem in Geology

In some studies the geologist can collect samples without restriction. Here are included studies of present-day environments of sedimentation, inasmuch as samples of the deposits may be collected from anywhere in the environment. Geomorphology is outstanding in having this advantage, inasmuch as the landscape features of interest are directly available for measurement in the field, from aerial photographs, or in many instances from large-scale topographic sheets.

Most studies of rocks and fossils are subject to more or less severe sampling restrictions because of superficial coverings of soil, of slumped deposits along valley walls, or because of coverings of glacial drift.

Such out-crops as occur may be controlled by differences in resistanc of the underlying bedrock, so that the exposures may have systemati rather than random errors. Samples collected from the outcrop area may therefore be biased with respect to some properties of the unex posed portions of the rock body.

Partly because of limitations on sampling, plus the absence half a century ago of principles of sampling adaptable to such situations geologists early developed the practice of examining all outcrop exposing the rocks of interest, and of then selecting purposively one o more 'typical' rock specimens, or sets of 'typical' fossils, from th formations under study. Measurements and observations on thes samples were summarised, sometimes as descriptive statistics, with th geologist assuming full responsibility for any generalisations that h extended from his samples to the complete rock body, including all o part of its buried extent.

Although this procedure may give the impression that geologica observations are invariably biased, it would be a serious error to accep this conclusion. Geologists have since the beginnings of their scienc regularly applied rigorous though qualitative principles of stratificatio to their informal geological populations, and have been careful t extend their generalisations only to the strata involved. (It is note worthy that 'stratified sampling' as practised by statisticians is really a formalisation of a basic geological technique, including even its name. Purposive selection of samples common in some geological studies such as collection of rock samples near the margins rather than withi rock bodies, tends to overemphasise extreme variants rather than majo bulk attributes of the rock.

It is true that bias may enter even the most carefully collecte specimens when the whole population of interest is not available fo observation and sampling. It is also true that generalisations or pre dictions derived from purposively selected samples cannot be expresse objectively in terms of the likelihood that they are correct withi specified limits. Fortunately, recent advances in sampling theory hav provided means for more objective sampling procedures when onl part of the population of interest is accessible.

In 1954 Cochran, Mosteller, and Tukey[3] divided populations int two broad types: a 'target population' that is the object of interest an about which inferences are to be made from sample observations; an an associated 'sampled population' representing those parts of the targe population from which the actual samples are drawn. Each elemen in the sampled population has some chance of entering the final sample When all elements of the target population are directly available fo sampling, the sampled population is identical with the target population

In geological terms the target population may be represented by al the sand grains in a layer of sandstone. The sampled population i represented by all the sand grains in the exposed portions of the roc layer. Although the geologist has no control over the natural occurrenc

of outcrops, once on an outcrop he has complete freedom in sampling the exposed rock layer. Thus, in the terms of Cochran, Mosteller, and Tukey, the geologist may collect 'lower semiprobability' samples in the sense that although he has no control of the upper level of sampling, he does have control of the lower level. Inferences can in general be made on a rigorous statistical basis from sample to sampled population, although the geologist himself must assume responsibility for any extensions he makes of these inferences to the target population.

Geologists were quick to grasp the implications of target and sampled populations. Rosenfeld[4] applied the concept to his study of a Lower Palaeozoic sandstone over the region New York to Kentucky, and he argued from the geological evidence that inferences made about the sampled population could be applied without serious bias to the target population. Griffiths[5] set the general concept into a geological framework by defining the 'hypothetical population' as a volume of rock that was actually at one time present in some part of the earth's crust, but that may have been in part removed by subsequent erosion. The missing parts obviously cannot be sampled. That part of the original rock body still extant is the 'existent population'. Of this only parts may be available for sampling, and these parts constitute the 'available population'.

It is evident that Cochran, Mosteller, and Tukey have unlocked an important door that gives geologists access to new ways of assessing their difficult sampling problems. It is a clear illustration of the contributions of a methodological science to a substantive science.

With respect to those statistical studies that do not have important sampling restrictions, geologists have increasingly designed their projects in terms of conventional sampling plans, especially during the 1950's. Cochran's book on sampling[6] gave much impetus to this development. An increasing tendency for use of a variety of randomisation procedures in several stages of geological studies (as in sampling itself, in assigning samples to some order of analysis, in selection of individual grains for measurement under the microscope, and so on) is readily apparent in current geological statistical literature.

The Problem of Too Many Variables

As mentioned earlier, even the seemingly simplest geological study involves a complex of interacting factors that need to be 'sorted out' by the geologist. In his qualitative evaluation of multivariate phenomena the geologist has learned to weigh his observational data by comparison and contrast, selecting from the innumerable details those underlying factors that appear to have controlled a given process. The internal consistency of the data and the implications of several lines of evidence simultaneously considered commonly provide a basis for selecting some single set of conditions that most satisfactorily accounts for the phenomenon. Chamberlin[7] years ago described the method of 'multiple

working hypotheses' in what has become a classical paper on geological methods of reasoning.

It is an interesting paradox that as numerical data entered geology some of the qualitative multivariate aspects of geological analysis had to be reframed as quantitative univariate parts. When moderate amounts of numerical data on attributes of present-day sedimentary deposits began to be obtained—forty years ago, perhaps?—the problem of treating half a dozen variables simultaneously by statistical methods was hardly feasible. Instead, the geologist 'sorted out' his variables qualitatively, and selected one or two that were obviously important for more detailed analysis.

As a result of this fragmentation of some complex geological problems, statistical studies tended to centre on particular aspects of larger problems. In the study of sedimentary deposits, for example, the particle size distributions of the sediments and the mineralogical composition of the particles were early selected for quantitative studies: to relate particle size to the geological factors that control this attribute and to use mineral composition as a guide to the parent rock and geographic area of origin of the deposit.

During the past decade geologists have attempted to integrate the results of specialised studies into more general frameworks of analysis. Griffiths[8] for example has considered the question of those fundamental properties of sediments that are necessary and sufficient to define a rock specimen, and which may be of major importance in the analysis and interpretation of ancient sedimentary deposits.

The problem of analysing the phenomena of sedimentation necessarily includes consideration of the processes by which the sediments are formed. If extemporaneous remarks are permissible here, one may define some aspects of the problem by an implicit function of the following nature:

$$f[p, G, P, (U, V, W), T] = 0 \qquad \ldots . (1$$

where $p_1, p_2, \ldots$ represent an unspecified number of physical, chemical and mineralogical properties of the sediments; $G_1, G_2, \ldots$ represent a number of geometrical and other properties of the deposit as a whole (such as the slope of a beach foreshore, the width of an offshore bar); $P_1, P_2, \ldots$ represent individual elements of geological processes, such as wave characteristics on a beach, current velocity in a stream, and so on; (U, V, W) represent geographic coordinates including elevation W above or below sea level; and T is a time or stratigraphic factor. Additional factors in the problem, such as biological controls on sedimentation and the fossil content of the deposits, could be included.

It seems doubtful whether the general form of equation (1) is capable of solution, inasmuch as many of the variables that would need to be included are not yet defined in terms of operational definitions for numerical measurements. From this implicit function, however, may be selected a variety of specific problems for study:

$$p_1 = f_1(p_2, p_3, \ldots) \quad \ldots . (2)$$
$$p_1 = f_2(U, V) \quad \ldots . (3)$$
$$(p_1, p_2, \ldots) = f_3(P_1, P_2, \ldots) \quad \ldots . (4)$$
$$(P_1, P_2, \ldots) = f_4(U, V, W) \quad \ldots . (5)$$
$$P_1 = f_5(T) \quad \ldots . (6)$$

Here equation (2) represents the study of a particular sediment property s a function of other properties and equation (3) is a study of its areal ariations over some area of interest. Equation (4) is a study of the ependence of several sedimentary properties on selected geological rocess elements; equation (5) is a study of areal variations in geological rocess elements over a sedimentary environment, such as changes in vave characteristics due to refraction within a bay; and equation (6) epresents a study of the change through time of some process element, uch as discharge of a stream.

The advent of high-speed computers and improvements in geological neasuring techniques provide a basis for returning to the fundamental nultivariate nature of geological analysis, this time on a quantitative asis. Formal multivariate methods have been applied,[9, 10] including se of discriminant functions.[11, 12] Multiple regression analysis, not ecessarily wholly linear, offers possibilities for 'sorting out' the relative nportance of variables in complex problems. I have recently[13] examned the relations among several properties of beach sand by regression nalysis, as in equation (2). Factorial designs, as pointed out by riffiths,[8] also lend themselves well to study of inter-relationships mong particle properties.

In an empirical stage of analysis the reduction in the sum of squares f a dependent variable is an important criterion of the relative imortance of the independent variables. Inasmuch as this reduction neasures a degree of mathematical association, and not necessarily a ausal relation, considerable judgement is required in selecting geogically meaningful variables for analysis. Here again the intuitive udgement of the geologist comes in strongly.

The Problem of a 'High Noise Level'

One reason for exercising considerable judgement in applying the um of squares criterion in geology is that many geological data have a rge natural variability owing to simultaneous activity of numerous ctors operating on different time and space scales. These sources of ariability may enter the data at any sampling level, and in some udies they tend to obscure underlying relations by their presence. hey seem also to contribute occasionally to spurious associations mong variables, and hence need to be taken into account where ossible.

Small-scale variations in geological data rise to major importance in udies of areal variations in geological phenomena, and they are well own on contour-type maps prepared by drawing lines of equal value rough a geographic distribution of sampling points. At each location

measurements may be made on sedimentary properties, on distribution of organisms in the bottom muds, and on physical or chemical conditions of the environment, such as the salinity of water in a tidal lagoon. The contour patterns on a series of such maps (usually only one variable is shown on a single map) are compared to gain information about associations among mapped variables in terms of the similarities or differences among trends that may be present.

Contour-type maps have in fact contributed much to the interpretation of the effects of several variables operating simultaneously over an area of interest, and they provide another way for geologists to examine the multivariate aspects of their science. Many such maps, showing areal variations in sediment properties or in various attributes of stratigraphic units (facies maps) have appeared in the geological literature. These are studies based on equation (3) of the previous section. Good examples of sedimentary maps readily available to British readers are in Allen's[14] study of the Top Ashdown Pebble Bed of Lower Cretaceous age in Sussex, Kent, and Surrey. Goodlet[15] presents several facies maps of the Lower Limestone Series of Carboniferous age in the Midland Valley of Scotland. Some of Allen's maps have locally 'spotty' patterns that represent small-scale components superimposed on larger-scaled systematic effects. Goodlet's maps show fairly regular patterns, though the effects of smaller-scaled variations are apparent here and there.

Trend surface analysis of maps, which consists essentially in fitting polynomial surfaces to the map data, affords one way of separating the small-scaled variations from the more systematic patterns in the mapped data. A number of writers have looked into trend surface analysis of geophysical data,[16] facies data,[17] sedimentary data,[18] and variations in igneous rocks.[19] I was initially brought into contact with this general approach by the work of Box[20] on response surface analysis.

The point to be emphasised in this discussion is that these small-scale fluctuations are largely inherent in the data, and are only in small part introduced by measurement error. Measurement error in well designed maps (as in the cited examples) seldom accounts for more than 10% of the total variability in the data, whereas the deviations from trend surfaces fitted to the maps may sometimes contain 50% or more of the total map variability.

Large residual variations, as in map residuals, in the 'within-group' variations of factorial experiments, or in deviations from regression functions in general, complicate the task of identifying important main effects, and make it more difficult to establish differences between classes of data. On the estimation side of statistics, such large residual variabilities also widen confidence bands about predictions and may imply to non-geologists that the investigation is not of the best quality if so much uncertainty is present in the data.

There is some tendency to apply the term 'noise' to the smaller-scale variations in geology. (In a geological sense, 'noise' may be applied to those fluctuations in data that cannot be assigned to specific causes

nd which, if they are large, may obscure the meaningful information n the data.) Despite this tendency, and the caption at the head of this ection, it is likely that much of the residual variability in geological tudies actually does carry important geological information. Map esiduals, though treated as random components by least-squares nethods of analysis, do in many instances bear a relation to geological vents. As an example, residuals in gravity maps are sometimes closely elated to occurrence of ore bodies, as Grant[21] shows. In facies maps ne residuals on trend surfaces are sometimes more important in their elations to oil occurrence than are the trend surfaces themselves.[17]

A major problem facing geologists in contour-type map interpretaon is identification of the 'geological signals' contained in the residual noise'. Trend surface analysis offers considerable promise, and it is kely that functions less empirical than polynomials will also be used. n this work high-speed computers are essential, and the task is one of stablishing criteria for detection of 'secondary trend components' in he data. Fortunately in many maps the residual deviations from the nain trend are essentially normally distributed, which favours their use or stating the reliability of predictions made from the trend. Study of requency distributions of mappable variables is facilitated by removal f the trend, but I cannot say whether normally distributed residuals rom the main trend will be skewed when secondary trend components re removed.

ome Concluding Remarks

I have emphasised three general problems of statistical analysis in eology that are close to my own work. It would be incorrect to say hat these are the only problems or even that they are the most impor ant problems, although sampling is obviously basic to any statistical tudy in geology. It would probably be best to consider the foregoing iscussion as a somewhat biased sample of some statistical activities in eology during the past ten years or so.

In a broad way, though perhaps grossly oversimplified, the progress f statistical thinking in geology from description to analysis can be ummarised by three statements regarding the mean particle size of and grains on some beach of interest:

(1) The mean particle size is 0·43 mm.
(2) The mean particle size is $0{\cdot}43 \pm 0{\cdot}06$ mm.
(3) The interval 0·30 to 0·56 mm has probability 0·95 of containing the 'true' mean particle size.

n the first of these the geologist merely states as a fact what he observed y detailed measurements on one or more samples of beach sand, erhaps purposively selected as 'typical samples'. In the second state ent he gives his reader information on the variation observed among is samples, perhaps on a straight least-squares basis. The third state ent has specific scientific meaning in that it presents a statistical

generalisation accompanied by the likelihood that it is correct within specified limits.

It is evident that the largest gap occurs between the second and third statements, and for rigorous application of the third statement to the beach the geologist needs to start with a target population, he needs to collect probability samples in accordance with some underlying sampling plan, he needs to look into the frequency distributions with which he is dealing, he needs to make his measurements on an operational basis, and in general he needs to include a great deal of geological reasoning to make certain that he has properly defined in his own mind the geological objectives of his study.

Some present-day aspects of geological analysis are in the first or second stages, owing to sampling or other difficulties as discussed above. There is clear evidence, however, that geologists are moving increasingly to the third stage as they overcome some of the problems of applying modern probability theory and statistical analysis in their science. In this progress statisticians can be of increasing aid as they become more acquainted with statistical activities in geology.

Acknowledgements

The writer is indebted to numerous geologists and statisticians for opportunities to discuss various aspects of the material presented here. During his stay in Great Britain he has been hospitably received at a number of universities where parts of this material were presented, and he is particularly indebted to Dr J. F. Kirkaldy, Chairman of the Department of Geology at Queen Mary College, for providing office facilities and extending many courtesies.

REFERENCES

[1] Fisher, Sir Ronald A. (1953). 'The expansion of statistics', *J. R. Statist. Soc.*, A **116**, 1.

[2] Krumbein, W. C. (1960). 'The "geological population" as a framework for analysing numerical data in geology', *Liverpool and Manchester Geol. J.*, **2**, 341.

[3] Cochran, W. G., Mosteller, F. and Tukey, J. W. (1954). 'Principles of sampling', *J. Amer. Statist. Assoc.*, **49**, 13.

[4] Rosenfeld, M. A. (1954). 'Petrographic variation in the Oriskany sandstone' Ph.D. thesis, Pennsylvania State University. Abstract, *Bull. Geol. Soc. Amer.* **65**, 1928.

[5] Griffiths, J. C. (1960). 'Sampling sediments for petrographic analysis', Chapter 16 in Milner, H. B., *Sedimentary Petrography*, Vol. 1. Allen and Unwin, London (in press).

[6] Cochran, W. G. (1953). *Sampling Techniques*. John Wiley and Sons, New York.

[7] Chamberlin, T. C. (1897). 'The method of multiple working hypotheses', *J. Geol.*, **5**, 837.

[8] Griffiths, J. C. (1960). 'Measurements of the properties of sediments', *Revue de Geog. Phys. et de Geol. Dynam., Paris* (in press).

[9] Burma, B. H. (1949). 'Studies in quantitative paleontology, II. Multivariate analysis', *J. Paleont.*, **23**, 95.

Dawson, K. R. (1958). 'An application of multivariate variance analysis to mineralogical variation, Preissac-Lacorne batholith, Abitibi County, Quebec', *Canadian Mineralogist*, **6**, 222. For the theory of the crossed design used here, see Krumbein, W. C. and Tukey, J. W. (1956). 'Multivariate analysis of mineralogic, lithologic, and chemical composition of rock bodies', *J. Sediment. Petrol.*, **26**, 322.

Emery, J. R. and Griffiths, J. C. (1954). 'Differentiation of oil-bearing from barren sediments by quantitative petrographic analysis', *Producers' Monthly*, **19**, 33.

Miller, R. L. (1954). 'A model for the analysis of environments of sedimentation', *J. Geol.*, **62**, 108.

Krumbein, W. C. (1959). 'The "sorting out" of geological variables illustrated by regression analysis of factors controlling beach firmness', *J. Sediment. Petrol.*, **29**, 575.

Allen, P. (1949). 'Wealden petrology: the Top Ashdown Pebble Bed and the Top Ashdown sandstone', *Quart. J. Geol. Soc. Lond.*, **104**, Part 2, 257.
See also Allen, P. (1959). 'The Wealden environment: Anglo-Paris Basin', *Phil. Trans. Roy. Soc.*, B, **242**, 283.

Goodlet, G. A. (1957). 'Lithological variation in the Lower Limestone Group in the Midland Valley of Scotland', *Bull. Geol. Surv. G. B.*, No. 12, 52.

Oldham, C. H. G. and Sutherland, D. B. (1955). 'Orthogonal polynomials: their use in estimating the regional effect', *Geophysics*, **20**, 295.

Krumbein, W. C. (1959). 'Trend surface analysis of contour-type maps with irregular control-point spacing', *J. Geophys. Res.*, **64**, 823.

Miller, R. L. (1956). 'Trend surfaces: their application to analysis and description of environments of sedimentation', *J. Geol.*, **64**, 425.

Whitten, E. H. T. (1959). 'Composition trends in a granite: modal variation and ghost stratigraphy in part of the Donegal granite, Eire', *J. Geophys. Res.*, **64**, 835.

Box, G. E. P. (1954). 'The exploration and exploitation of response surfaces: some general considerations and examples', *Biometrics*, **10**, 16.

Grant, F. (1957). 'A problem in the analysis of geophysical data', *Geophysics*, **22**, 309.

THE CONSISTENCY OF SETTING OF A MACHINE TOOL HANDWHEEL

E. NIGEL CORLETT and GEOFFREY GREGORY*
Department of Engineering Production, University of Birmingham

An experiment was carried out to determine the effect of certain factors on the consistency of setting of a machine tool handwheel. A half-replicate of a 2^7 factorial design was used, confounded into 4 blocks of 16 elements. The design, analysis, and interpretation of the experiment are described in this article.

Introduction

It is only recently that machine design has been considered in relation to the operator, one of the incentives towards this approach having been the necessity to operate complex equipment in the armed services during the last war. It is still not unusual for equipment to be built and then for selection procedures to be found to be necessary in order to find suitable operators. Fortunately, this approach is on the decline and the more rational approach of designing equipment and controls suitable for operation by human beings is being introduced.

In order to determine what is the most suitable arrangement of, or working conditions for, the controls of military equipment, many experiments have been carried out; amongst them have been some to decide the size, position, and resistance of handwheels in order to give the most consistent performance in gun laying or radar tracking.[1] These experiments indicated fairly well-defined optima, suggesting that an investigation of the various factors involved in setting a machine tool handwheel might similarly produce results which would give guidance for designers of machine tools requiring certain levels of accuracy or consistency.

The factors to be taken into account were listed, and apparatus as shown in Plate I was then built. It consisted of an optical dividing head (A) with a dial (B) mounted directly on to its mandrel. Connected to the mandrel by a universal coupling was the handwheel spindle, provided with an adjustable friction brake (C). Provision was made for changing both dials and handwheels. A framework with a removable platform provided variations in floor level and was capable of adjustment by inches over a range of 15 in.

By experiments it was desired to find the consistency of setting of operators using this apparatus. To determine the factors which would require detailed investigation, using several subjects, it was desirable that a survey experiment should be conducted covering all the possible factors which were likely to affect the performance. In order to

* Dr Gregory is now Senior Lecturer in Statistics at the University of Melbourne.

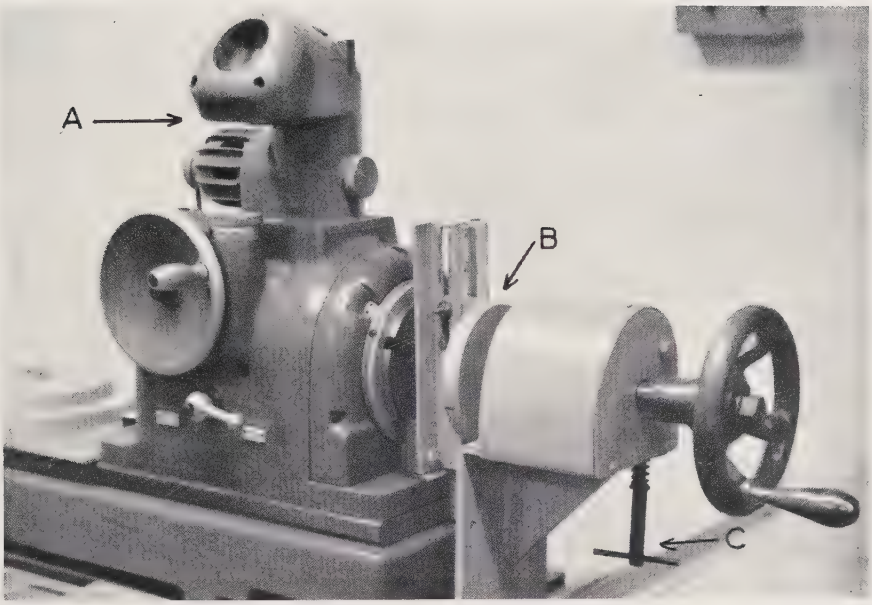


PLATE I. The Apparatus.

determine 'consistency' it was not necessary to determine the variation of the settings from an absolute value of the origin, since whether the operator sets consistently 'high' or 'low' is immaterial in the industrial situation, where he knows from experience the effect of his setting and adjusts himself accordingly. What is important is that his adjustment to his own peculiarity should also be consistent. Thus the design of the apparatus, the experiment, and also the criterion of measurement must be such that the effects of the factors on consistency can be determined. No 'absolute zero' for each dial was required, the variance of the readings for each run being a measure of the consistency. It was conjectured that a sample of 25 readings was the minimum possible number to give a reliable estimate of the variance.

Preliminary runs determined the way the experiments would be conducted. Each time an observation was taken the subject first off-set the dial through an angle of approximately 15°. He then brought the line of the dial into coincidence with the fixed line on the dividing head, making the final adjustment by means of a series of taps by hand on the handwheel rim. The approach was always made in a clockwise direction, and overshooting the line necessitated a complete reversal and a new approach. The whole operation took between 5 seconds and about a minute, depending on the difficulty of the operation. Subjects were told to aim for as high and consistent a degree of accuracy as they could throughout the experiment. Recording of the reading took another few seconds, and changing the conditions of the apparatus between 5 and 10 minutes. Thus each turn required about 15 minutes to complete, less than half of which was operating time.

The choice of factors to be tested was based on the conditions existing in the industrial situation. Illumination and contrast between line and dial were held at values recommended by advanced practice, and the spindle axis was kept horizontal at right angles to the operator for this series of tests.

Factors which were varied were:

(i) handwheel diameter (H),
(ii) dial diameter (D),
(iii) line thickness (T),
(iv) spindle friction (F),
(v) working height (L), taken as the height of the operator's elbow, with his arm by his side, relative to the height of the centre of the handwheel.

It was felt that these were the basic factors of the problem. With each of the factors at two levels a complete 2^5 factorial experiment would take 8 hours to complete. Two levels of each factor were in fact selected, the levels being approximately at opposite quartiles of the accepted practical range. A detailed investigation of the response surfaces of the more significant factors is planned for subsequent experiments.

The actual levels selected were:

H $5\frac{1}{2}$″ diameter and 10″ diameter.
D 4″ diameter and 8″ diameter.
T 0·008″ thickness and 0·064″ thickness.
F $7\frac{1}{2}$ lb.-in., and 45 lb.-in.
L elbow level with centre line of spindle, and 6″ above centre line of spindle.

The experiment was still too large to be accomplished by one subject without causing loss of concentration due to fatigue or eye-strain. At the expense of confounding high-order interactions, the experiment could be performed in sessions of a more reasonable duration. It was also realised that other factors might be introduced, which, although perhaps not of the importance of the original five, could be evaluated at the cost of very little extra experimentation and loss of information. Two factors were therefore introduced.

(vi) Practice of the operator (P). The experiment was performed by two subjects, one of whom was an experienced craft apprentice, the other being entirely unaccustomed to the use of this type of machinery. Differences in the performances of the two subjects have in fact been attributed to the difference in their experience. The subjects were similar in height and eyesight but it is recognised that other features of their history or physique may have influenced their performances even more than the obvious difference of practice with the particular type of machine tool.

(vii) Knowledge of results (K). In some machine situations the operator obtains knowledge of the accuracy of his setting by measurement of the workpiece. This was simulated by telling the subject the reading of the dividing head immediately after each setting.

The Design

A complete factorial experiment would now have consisted of 128 factor combinations.[2] In a morning or afternoon session of four hours, sixteen runs could be performed, and this formed a natural block for the design. The amount of time each subject spent on the experiment was halved by confounding the main effect of practice (P) with the blocks. The main effect of practice on consistency of setting was no direct concern of the experiment, but on the other hand the interactions of P with the other factors were of interest and were preserved.

Experience suggested that first- and second-order interactions were likely to be significant. A half replicate with alias subgroup [(1), −HDTFLPK] was used.

In order to split the experiment into blocks of size 16, a confounding subgroup with two generators was used. One generator was the element P, and the second-order interaction HTK was selected as the other; a design was thereby achieved in which only two second-order

interactions HTK and DFL were confounded. These two were considered unlikely to have any reasonable interpretation.

The confounding subgroup was (1, P, HTK, PHTK), with alias (HDTFLPK, HDTFLK, DFLP, DFL).

Thus the 2^7 experiment was reduced to a half replicate confounded into 4 blocks of 16 elements.

The model used in the analysis assumed that the treatments have a multiplicative effect on the variance. This was partly because it seems more reasonable than an additive effect, and partly because it would improve the homogeneity of the residual variability, as well as simplifying the analysis. Using log (s^2), where

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

and $x_1, x_2, \ldots . x_n$, are the observed values, the distribution of log (s^2) has approximately a constant variance irrespective of the distribution of x:

$$\text{var}[\log_e (s^2)] = \frac{2}{n-1} + \frac{\gamma_2}{n} + O\left(\frac{1}{n^2}\right)$$

where γ_2 is related to the kurtosis of the x-distribution, and O signifies 'of the order of'. For near-normal distributions the last term may be neglected, and if γ_2 is sufficiently small the variance is seen to depend only on n. For normal distributions of x, γ_2 will be zero.

Confidence intervals for the estimate of kurtosis

$$b_2 = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2}$$

under the assumption of normality are not, to our knowledge, tabulated for a value of n as small as 25. However, using an alternative estimate of kurtosis we can test the hypothesis that the kurtosis is consistent with a normal distribution of the individual readings. The statistic is that suggested by R. C. Geary:[4]

$$a = \frac{\sum_{i=1}^{n} | x_i - \bar{x} |}{\left[n \sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{\frac{1}{2}}}$$

If this statistic is consistent with normality, it is reasonable to take γ_2 as zero in the above equation.

The design (before randomisation within the blocks) and the results, $\log_{10} (s^2)$, are shown in Table I. The presence of a letter in an element corresponds to the factor at its 'larger' level—e.g. larger handwheel, higher working height, greater friction, practised operator, operator with knowledge, etc.

TABLE I

Design and results

Block 1		Block 2	
(1)	0·7044	hd	0·0561
ht	0·5907	dt	0·3615
tk	−0·0297	hdtk	−0·1158
hk	0·3914	dk	−0·1952
df	0·0792	hf	0·4585
htdf	0·3228	tf	0·2531
tkdf	−0·1599	htkf	0·2727
hkdf	−0·0996	kf	0·6861
fl	0·5878	hdfl	−0·0074
htfl	0·3577	dtfl	−0·2328
tkfl	0·1847	hdtkfl	−0·1046
hkfl	0·5706	dkfl	−0·2069
dl	−0·1805	hl	0·8051
htdl	−0·3224	tl	0·4634
tkdl	−0·1433	htkl	0·2904
hkdl	0·1354	kl	0·4692
Block 3		**Block 4**	
pd	−0·6760	ph	0·5457
pdht	−0·3824	pt	0·0846
pdtk	−0·2996	phtk	0·4453
pdhk	−0·4539	pk	0·2361
pf	0·2970	phdf	−0·5069
phtf	0·1646	ptdf	−0·3299
ptkf	0·3878	phtdkf	−0·3245
phkf	0·2168	pkdf	−0·3233
pdfl	0·0148	phfl	0·4199
pdhtfl	−0·4898	ptfl	0·2957
pdtkfl	0·1308	phtkfl	0·2278
pdhkfl	−0·1829	pkfl	0·4269
pl	0·0182	phdl	−0·4798
phtl	0·2070	ptdl	−0·0669
ptkl	0·1101	phtkdl	−0·0584
phkl	0·2642	pkdl	−0·6856

Analysis of the Experiment

The average of the 64 values of a computed from the readings within each run was 0·808, in close agreement with the population mean of 0·806 when the distributions are normal.[3] Of the 64 values, one lay outside the 2% significance limits, six lay outside the 10% significance limits, and nine lay outside the 20% significance limits. Thus it seemed fair to assume that the residual errors from the assumed model have a distribution with constant variance equal to

$$(\log_{10} e)^2\left(\frac{2}{n-1}\right) = 0{\cdot}0157.$$ The standard deviation $= 0{\cdot}125$.

The differences from zero of the estimates of the main effects and interactions were compared directly with their standard deviation. Each estimate is the sum or difference of 64 observations. Their standard deviation is therefore 0·0156. Although the distribution of residuals may not be precisely normal, it is reasonable to assume normality in testing the differences from zero of the estimates, since the latter are based on so many observations.

To be significant at the 0·1% level, a main effect or interaction

would need to be greater than 0·0513 in absolute value, and the corresponding values for the 1% and 5% levels of significance are 0·0402 and 0·0306 respectively.

The analysis of the data is shown in Table II.

TABLE II

Estimates of factors and interactions

Factor and Alias		Estimate	Factor and Alias		Estimate
(1)	– HDTFLPK	0·0856†	K	– HDTFLP	– 0·0211
H	– DTFLPK	0·0149	KH	– DTFLP	0·0128
T	– HDFLPK	– 0·0202	KT	– HDFLP	0·0066
HT	– DFLPK	– 0·0126	KHT	– DFLP	0·0132†
D	– HTFLPK	– 0·2708***	KD	– HTFLP	0·0134
HD	– TLFPK	– 0·0180	KHD	– TFLP	0·0327*
TD	– HFLPK	0·0670***	KTD	– HFLP	0·0051
HTD	– FLPK	– 0·0301	KHTD	– FLP	– 0·0292
L	– HDTFPK	0·0025	KL	– HDTFP	0·0223
LH	– DTFPK	– 0·0009	KLH	– DTFP	0·0267
LT	– HDFPK	– 0·0148	KLT	– HDFP	0·0188
LHT	– DFPK	– 0·0410**	KLHT	– DFP	– 0·0040
LD	– HTFPK	0·0027	KLD	– HTFP	0·0260
LHD	– TFPK	– 0·0047	KLHD	– TFP	0·0232
LTD	– HFPK	– 0·0129	KLTD	– HFP	0·0459**
LHTD	– FPK	0·0096	KLHTD	– FP	– 0·0302
P	– HDTFLK	– 0·1095†	PK	– HDTFL	0·0524***
PH	– DTFLK	– 0·0151	PKH	– DTFL	– 0·0031
PT	– HDFLK	0·0506**	PKT	– HDFL	0·0331*
PHT	– DFLK	– 0·0198	PKHT	– DFL	0·0049†
PD	– HTFLK	– 0·0249	PKD	– HTFL	0·0003
PHD	– TFLK	– 0·0219	PKHD	– TFL	0·0175
PTD	– HFLK	– 0·0053	PKDT	– HFL	– 0·0002
PHDT	– FLK	0·0166	PKHTD	– FL	– 0·0162
PL	– HDTFK	0·0309*	PKL	– HDTF	– 0·0340*
PLH	– DTFK	– 0·0198	PKLH	– DTF	0·0181
PLT	– HDFK	0·0195	PKLT	– HDF	– 0·0201
PLHT	– DFK	0·0215	PKLHT	– DF	– 0·0137
PLD	– HTFK	0·0563***	PKLD	– HTF	– 0·0311*
PLHD	– TFK	– 0·0099	PKLHD	– TF	0·0259
PLTD	– HFK	0·0223	PKLTD	– HF	0·0397*
PLHTD	– FK	– 0·0217	PKLHTD	– F	– 0·0203

† Confounded with blocks
* Significant at 5% level
** Significant at 1% level
*** Significant at 0·1% level

The most significant effect is that of the dial diameter. This is hardly surprising since the error measured is an angular error. There is also a highly significant interaction with the thickness of the line on the dial. The data are summarised in Table III.

TABLE III

Mean values of log (s^2)

	4″ Dial	8″ Dial
0·008″ Line	0·444	– 0·232
0·064″ Line	0·269	– 0·138

S.D. 0·031

The figures quoted are averages of 16 observations and each therefore has a standard deviation of 0·031. The larger dial results in considerably more consistent settings, particularly when the narrower line

is used. Further experiments investigating the relationship between dial size and line thickness are planned.

The interaction PLD is highly significant and also PL is significant at the 5% level. Analysis of the results is shown in Tables IV and V.

TABLE IV

Mean values of log (s^2)

	Floor Level: Low	High
Subject with no practice	0·224	0·167
Subject with practice	−0·057	0·009

S.D. 0·031

TABLE V

Mean values of log (s^2)

		4″ Dial	8″ Dial
Subject with no practice	Low floor level	0·416	0·031
	High floor level	0·466	−0·133
Subject with practice	Low floor level	0·297	−0·412
	High floor level	0·246	−0·227

S.D. 0·044

The results of changing the floor level are interesting. It is possible that they may be affected by individual differences between the subjects and it would not be possible to determine this, of course, without further experiment. (The same applies to the other interactions involving P.) However in what follows we shall take the results at their face value as indications of promising directions for further work. Considering the results using either dial, there is some evidence that the unpractised subject prefers the higher floor level, whereas the practised subject obtains better results at the lower floor level. The effect of the PLD interaction is shown clearly in Table V, where the evidence is apparent when both subjects use the 8″ dial, but it is reversed and less definite when the 4″ dial is used. If, then, further experiments are to be carried out on various dials and thicknesses, and if the dials used are of diameter around 8″, it seems that the best results will be obtained with a subject who has had some practice if he operates at the lower floor level. If, however, a subject who has had no practice is to be tested, then there is evidence that more consistent results will be obtained if the floor level is higher.

The remaining interaction significant at the 0·1% level is the interaction PK. The inference from this is that the more experienced subject was distracted by the knowledge of the results of each setting, or equivalently that the less experienced subject improved this consistency through knowledge of his results. Table VI gives the averages.

TABLE VI

Mean values of log (s^2)

	No Knowledge	Knowledge
Subject with no practice	0·269	0·122
Subject with practice	−0·055	0·007

S.D. 0·031

The interaction PT is significant at the 1% level (Table VII) and also PKT is significant at the 5% level (Table VIII).

TABLE VII

Mean values of log (s^2)

	0·008″ Line	0·064″ Line
Subject with no practice	0·266	0·124
Subject with practice	−0·054	0·006

S.D. 0·031

Thus the practised subject appears to be more consistent with the thinner line, whereas the reverse is true for the unpractised subject. However, when practice, knowledge, and thickness of line are considered together, the interaction PKT becomes apparent (Table VIII).

TABLE VIII

Mean values of log (s^2)

	No Knowledge		Knowledge	
	0·008″ Line	0·064″ Line	0·008″ Line	0·064″ Line
Subject with no practice	0·313	0·224	0·219	0·024
Subject with practice	−0·046	−0·065	−0·063	0·077

S.D. 0·044

Thus for the practised subject the combination of no knowledge of the results of his settings and the thinner line does not yield the most consistent accuracy of setting (three of the four results are, of course, not significantly different). If in subsequent experiments the thickness of the line is to be varied over the range considered in the preliminary experiment, the most consistent results seem to be obtained by telling an inexperienced subject of his results, but withholding the information from an experienced subject. (For the purpose of the present discussion, we are discounting the possibility that 0·077 is a freak result.) If, however, the thickness of the line is fixed in the region of 0·008″, then the

evidence suggests that it is better to inform all subjects of their results.

The interaction PHF is significant at the 1% level, and this coupled with the interaction HF significant at the 5% level suggests that the relationship between diameter of handwheel and the friction on the spindle may be a field for further experimentation.

The average effects of the larger handwheel and the greater amount of friction are summarised in Table IX.

TABLE IX

Mean values of log (s^2)

	Friction	
	7½ lb.-in.	45 lb.-in.
5½″ Handwheel	0·011	0·131
10″ Handwheel	0·120	0·081

S.D. 0·031

The inference from the data is that for the smaller handwheel less friction yields the more consistent results, but for the larger handwheel, the opposite is true, though the effect is not much greater than the S.D. The interaction HF is significant only at the 5% level, and more evidence would be needed before the response surface could be predicted with any accuracy.

When these results are split into those obtained from the two subjects, the following table is obtained (Table X).

TABLE X

Mean values of log (s^2)

		Friction	
		7½ lb.-in.	45 lb.-in.
Subject with no practice	5½″ Handwheel	0·181	0·149
	10″ Handwheel	0·229	0·221
Subject with practice	5½″ Handwheel	−0·160	0·112
	10″ Handwheel	0·011	−0·059

S.D. 0·044

The interaction between H and F of the previous table is now shown to be much more evident in the case of the experienced subject.

In the analysis each interaction significant at the 0·1% level has been considered and also all interactions significant at the 1% level except LHT. Interpretation of LHT is difficult since none of its component main effects or interactions is significant. Possibly this is a random effect. Certainly it would not be surprising if one out of the sixty-four effects were significant by chance at the 1% level. Of the six interactions significant at the 5% level, three have been interpreted with more significant interactions. There remain KHD, PKL and HTF. Again, it is plausible that three effects will be significant by chance at

the 5% level, which at this stage of the investigation is probably as good an explanation as any.

Conclusions

The most promising fields for further experiments appear to be:

(1) Vary dial diameter (D) and thickness of line (T).

(2) Vary handwheel diameter (H) and friction on the spindle (F).

In carrying out these experiments, the remaining factors will be held constant at values which tend to give more consistent results overall. Since decisions have to be made and the outcome of the decisions makes little difference to the complexity of the experiment, it is of no great consequence that some of the decisions are based on insufficient data.

Practice of the Operator (*P*). The more consistent results throughout the experiment have been obtained with the more experienced subject. The main effect P has, admittedly, been confounded with blocks, but there is no evidence of a block effect within subjects. Some estimate of the block to block variation is available from the other two confounded interactions. These in fact have a root mean square value of 0·00996, which is smaller than their theoretical root mean square. Thus there is apparently no significant effect of blocks. However, strictly speaking, the significance of P should be tested by a comparison with this estimate (0·01) of block variation (which, although subject to error, is the best we have), and this being based on only two degrees of freedom requires a high value of *t*. The observed value of *t* of 11 exceeds the 1·0% level of significance, (compare that the probability level is less than 0·1% if the theoretical root mean square is used) indicating a P effect. It is likely, therefore, that the significant effect of P confounded with blocks is in fact due to the difference in experience of the subjects.

Working Height (*L*). Tables IV and V indicate that if a practised subject is to be used, then the lower working height would give the better results. If in experiment 2 a large dial is being used, there is strong evidence for the lower level, and even for a 4″ diameter dial the preference for the higher level shown in Table V is not significant.

Knowledge of Results (*K*). Table VI suggests that more consistent results for a practised subject are obtained when he is not immediately informed of his results. Nevertheless, this effect fails to show with a line of thickness 0·008″.

Acknowledgements

This research has been carried out with the guidance and encouragement of Professor N. A. Dudley, Head of the Department of Engineering Production, University of Birmingham.

The authors also gratefully acknowledge the valuable assistance in the analysis of the experiment given by Mr M. H. Quenouille of the London School of Economics and Political Science, and Dr F. Benson of the University of Birmingham.

REFERENCES

[1] Hick, W. E. (1945). *Factors Affecting Handwheel Design.* Medical Research Council, Applied Psychology Unit, No. 26.

[2] Quenouille, M. H. (1953). *The Design and Analysis of Experiments.* Griffin, London.

[3] Pearson, E. S. and Hartley, H. O. (1956). *Biometrika Tables for Statisticians,* Vol. 1, p. 183. Cambridge University Press.

[4] Geary, R. C. (1935). 'The Ratio of the mean deviation to the standard deviation as a test of normality', *Biometrika,* **27**, 310.

STATISTICS AND TASTE TESTING*

NORMAN T. GRIDGEMAN
Division of Applied Biology, National Research Council, Ottawa, Canada

In this article Mr Gridgeman surveys some of the statistical problems that arise in the testing of taste. He deals with tests of discriminability and tests of preference, and discusses the question whether the subject should be allowed to record ties or should be compelled to give a definite vote even if he is not sure.

Introduction

Almost every sort of test comes into the purview of the statistician, whose direct usefulness however is a function of certain aspects of the experimental unit, such as its reliability, its representativeness, and its cost. When, for instance, the unit is biological, its representativeness in terms of a defined population is not obvious and the statistician can often prove his worth in the clarification process. No test is of interest *per se*; each has to be regarded as the middle section of a tripartite investigational entity, of which the first section is the theory on which the actual conduct of the test is based, and the last is the drawing of inferences from the results of the test. The theory may require a mathematical model; the testing may call for an experimental design; and the inference may be uncertain. These things are all in the domain of the statistician. He will usually be at his best in furnishing experimental designs, and at his most subtle (perhaps exasperatingly so) in making or circumscribing inferences. Among those specialties that can benefit from the statistician's insight, taste testing takes high rank.

Character of Taste

For present purposes taste may be loosely defined as a psychosensory reaction to chemical stimulation of certain areas of the tongue. Three of its facets are important. Two, quality and intensity, rest on objective characteristics; but the third, delectability, is subjective. Taste, obviously, is complex; moreover, it is rarely experienced in isolation. It is also less amenable to quantitative study than the senses of hearing and seeing.

The number of different kinds of taste is large; this statement holds true both for the theory that they are all composed of a few primaries (sweet, salt, sour, and bitter), and for the rival theory that countless unique sensations exist. We learn to identify many tastes, although we are perhaps less adept than we like to believe. Recognition is helped by such ancillaries as odour, appearance, and context, and if these are

* N.R.C. No. 5791. Based on a paper given at the Third Stevens Symposium on Statistical Methods in Industry, Hoboken, N.J., January 1959.

unusual we are easily misled. Some years ago Moir[19] conducted a classic experiment in which special batches of jelly were prepared with the colours and the corresponding flavours out of phase; this baffled all tasters.

Taste identification can be improved by learning, and some people can learn better than others. This is the key to the skill of the expert tasters. Additionally, experts have to make preferential (hedonic) decisions, which means that part of their skill must lie in having canons of acceptability and excellence and so forth that conform with those of the consuming public (or, more specifically, the relevant section of the public).

Taste and Measurement

Delectability and intensity are quantitative concepts, and this brings us sharply up against the problem of scaling. Manifestly, the *notion* of psychosensory continua for delectability and intensity is valid; the real question is whether they serve any useful purpose and whether we can associate scale points with them. I think that they do, and that we can. Not everyone will agree with this; nevertheless, as F. C. Bartlett[1] has remarked, there is 'almost complete agreement that, even if sensations cannot be measured, they must be'. Unless the principle of the measurability of a sensation is accepted, none but the most primitive sort of taste testing can be conceived. The available measures may be crude, but that is incidental. One way of scaling delectability in, for instance, frozen strawberries, is to describe a number of categories, to train a panel to agreement upon the interpretation of the descriptions, to rank the categories in a rational order, and then to assign numerical scores, in the same order, to the categories. A valid metric will then emerge. The same applies if several arbitrary degrees of a desirable characteristic, and/or an off-flavour, are specified. Even a simple labelling of categories as 'excellent', 'very good', and so forth, often serves, although the further we move away from agreed, well-defined classifications, the more nebulous and subjective and difficult to validate will the score scale be. One test of the interconsistency of scores is to check whether, for example, two batches whose mean scores are 2 and 6 can be mixed to form a batch whose mean score is 4. To carry out orthodox statistical analysis on the scores we need to check the assumptive requirements appropriate to the statistical model; fortunately the variances are usually found to be sufficiently homogeneous, and the distribution sufficiently close to normality, to justify the use of the analysis of variance.

More could be said about *absolute* scaling techniques, particularly when multiple criteria are being studied simultaneously (taste, odour, texture, etc.). Interesting controversies exist on the best methods of carrying out these tasks, and about whether they should be done in single-seater booths or by 'open appraisal' by the whole panel round a table; but these controversies are more technical than statistical, and

they will not be pursued here. Instead we shall move on to comparative work.

Comparative Tests

Comparative work is a major task of what is often called the analytical taste panel; and it lends itself well to probabilistic modelling. The devices most commonly employed are ranking or sorting or matching of samples according to some specified attribute, such as delectability or freshness. Pair comparison is a minimal case, and we shall use it as a foundation stone in the building of a rationale of comparative work in general. The scaffold will be the classical psychophysical structure—see, for instance, Cane[6] and Harper.[14] In particular, we shall consider the response of a single subject to a well-defined taste stimulus, say, sodium chloride.

The Model

Postulate (i) a sensory response-continuum for the intensity of sodium chloride, (ii) a monotonic function relating S, the concentration of the salt, to R, the corresponding point on the continuum, and (iii) a distribution of judgements, in replicate independent trials, of the location of a particular R. We further assume (provided that R is not near to the threshold of detectability) that this distribution is close to the normal (Gaussian) form, with R as the parameter of location. The other parameter (of dispersion) will be a measure of the subject's precision of judgement.

The question immediately arises: How can R be given a metric? This is a tough one, but, fortunately, it can be side-stepped. So long as we restrict our attention to binary ranking (pair comparison) in this setting, the problem of the R scale obligingly drops out of sight and significance. Consider a subject exposed to two different concentrations S_i and S_j (under code); then his response to the query, 'Which is saltier?' will be governed by his judgement of the interval between the quantities R_i and R_j. Imagine a series of such comparisons, with a common (fixed) S_j and various S_i. Then, granted certain statistical assumptions that need not be elaborated here, the probability that the answer will indicate S_i is

$$P(i>j) = \int_{-\infty}^{R_i - R_j} f(u)du \qquad \ldots . (1)$$

where $f(u)$ is some symmetrical near-normal function. An operationally convenient device is to make $f(u)$ the 'squared hyperbolic secant'—see, for example, Gridgeman[10] and Maxwell.[18] We then have

$$P(i>j) = \int_{-\infty}^{R_i - R_j} \tfrac{1}{4}\,\mathrm{sech}^2\frac{u}{2}\,.\,du$$

$$= \frac{1}{1+\exp(R_j - R_i)} \qquad \ldots . (2)$$

from which it follows that

$$\ln\left\{\frac{\mathrm{P}}{1-\mathrm{P}}\right\} = R_i - R_j \qquad \ldots\ldots (3)$$

Uses of Pair Comparison

The above relations underlie many taste experiments. Take, for instance, the man who is interested in assessing taste acuity, or who wishes to explore the function that connects stimulus and response. His first hypothesis might be the maligned Fechner relation,

$$R = \alpha + \beta \log S \qquad \ldots\ldots (4)$$

from which it follows that

$$R_i - R_j = \beta \log (S_i/S_j) \qquad \ldots\ldots (5)$$

Substitution of (5) in (3) gives a probabilistic model for the estimation of β, which is a parameter of discriminability functionally related to the *jnd* or 'just noticeable difference'. The only observations needed are frequency estimates of P, obtained from pair comparisons, for various S_i. The estimation problem is easy if we follow Berkson's treatment[2] of the analogous bioassay problem (this entails a non-iterative least-squares solution). A check of the fit of the data to the model presents no difficulties.

As another example, take the consumer surveyor who has $\mathcal{N}$ different varieties of a food product, and who would like to know whether they generate preference differences in a certain population. Having chosen a panel that fairly represents the population, he can ask the members to rank subsets of n ($\mathcal{N} \geqslant n \geqslant 2$) items in order of preference. There will be $\binom{\mathcal{N}}{n}$ such subsets, and $n!$ possible arrangements of each. When $n = \mathcal{N}$ the whole group is ranked in one operation—a task that is psychosensorily difficult if $\mathcal{N}$ is large. At the other extreme of $n = 2$ there are $\mathcal{N}(\mathcal{N}-1)/2$ subsets of pair comparisons, and Bradley and Terry[5] have devised a method of handling these based on equation (2). Using 'rankits', defined as the expected values of the deviates for the ranked items in a sample from a normal population with zero mean and unit standard deviation, Bliss[3] appropriately transforms preference frequencies derived from the $\mathcal{N}(\mathcal{N}-1)/2$ pair comparisons. This is a simple procedure that gives satisfactory results. Fisher and Yates's Table XX supplies the rankits.[9]

Other methods of dealing with multiple pair comparisons are associated with the names of Thurstone,[22] Mosteller,[20] Kendall,[16] Gulliksen,[13] Bose,[4] and Scheffé.[21] Informative discussions of these, as well as practical assessments of their relative merits, will be found in papers by Bliss *et al.*,[3] and by Jackson and Fleckenstein.[15]

The method of Scheffé needs special mention because it differs from the others in requiring not merely a straight choice, A or B, but a degree of preference. This departure from the binary system incurs some weakening of the statistical logic involved; nevertheless, there are com-

pensating advantages. More information is extracted. Scheffé recommends a 7- or 9-point scale; the former is constructed as follows:

Preference:	Score for A (=score for B with sign changed) *if for A*	*if for B*
strong	+3	−3
moderate	+2	−2
slight	+1	−1
none	0	0

When using the Scheffé method, I favour the elimination of the central zero point on the ground that its availability fosters indecision. The remaining categories must then be rescored: when, for instance, the 7-point scale is so reduced, the following scores, +5, +3, +1, −1, −3, −5, will preserve the requirement for 'equal-appearing intervals'. The scheme incorporates a test of the 'hypothesis of subtractivity', i.e. of the assumed existence of characteristic sample parameters, located on the hedonic continuum, that generate the preferences.

To Tie or Not to Tie

The question of a central zero point in hedonic scaling is part of the wider question of whether or not, in straight pair comparison, ties should be allowed. The answer is: it all depends. If we are concerned with testing discriminability, or if the null hypothesis that 'items A and B are indistinguishable' is being checked, ties are best avoided. The question put to the subject should be 'Which. . . .? Answer one or the other even if you feel you are guessing', rather than, 'Which, if either. . . .?' On the other hand, when preferences in a population are sought, the answer 'Neither' should be allowed. Some further remarks on this topic follow; they mostly stem from a recent investigation (Gridgeman[12]).

If the subject is judging the relative intensity of two taste stimuli, the test is purely of discrimination, for we may assume that, having detected the difference, he will not mistake its direction. When, on occasion, he does not detect the difference, he has to guess. Assume an empirical probability p_s of true discriminability. Then the left-hand side equation (1) can be written:

$$\mathrm{P}(i>j) = (1 \pm p_s)/2 \qquad \ldots . (6)$$

For a specific pair in which $C_i > C_j$, we have

$$\left.\begin{aligned} P_i &= \mathrm{P}(i>j) = (1+p_s)/2 \\ P_j &= \mathrm{P}(j>i) = (1-p_s)/2 \end{aligned}\right\} \qquad \ldots . (7)$$

Now if ties are allowed, the subject will not invariably record 'no apparent difference' when he fails to discriminate; sometimes (and not necessarily consciously) he will guess. We therefore postulate a probability p_g of his guessing (and, of course, about half of his guesses will be

correct). Then, writing P_0 for the probability of a tie, we have:

$$\left.\begin{aligned} P_i &= p_s + p_g(1-p_s)/2 \\ P_j &= p_g(1-p_s)/2 \\ P_0 &= 1-P_i-P_j = (1-p_g)(1-p_s) \end{aligned}\right\} \quad \ldots\ldots (8)$$

Equations (8) have as a special case $p_g = 1$, which means that guessing is obligatory, and which is represented by equations (7). In general, theory shows that ties are best banned in discrimination work *provided that* p_s *itself* (when non-zero) *is independent of the design.* But the design is rarely without effect on p_s. This is not surprising. Normally, the simpler the design, and the more straightforward the subject's task, the larger p_s will be—a circumstance that is apt to offset the theoretical advantages of the no-tie system.

Ties in Preference Tasting

When preference is at issue, three classes of subject can be distinguished:

Those (a proportion p_i) who prefer i
Those (a proportion p_j) who prefer j
The rest (who have no preference)

But not all of the third (the 'indifferent') class will say they have no preference; some of them, some of the time, will guess, and we assume that there is a probability p_g that any member of this class, on a particular occasion, will in fact state a pseudopreference. We must further suppose that such decisions will not be influenced by extraneous circumstances or characteristics of the compared items, so that there is an equal chance of choosing i or j in the absence of preference. To take an exaggerated example: it would be methodologically wrong to put chocolates 'i' in a handsome white pack embellished with a pretty girl, and to put 'j' in an ill-made red pack covered with a hammer-and-sickle motif. This does not however imply that all extraneous differences should be equalised; it may be best to retain them and 'balance them out'. Thus we might put half the 'i' chocolates in white packs and the other half in red packs, and ditto with the 'j' chocolates; then the two main pairings (white: red and red: white) or perhaps all four possible pairings (white: red and red: white, red: red, and white: white) should be equally distributed among trials. Then, ties being permitted, the composition of the three kinds of answer will be:

$$\left.\begin{aligned} P_i &= p_i + (1-p_i-p_j)p_g/2 \\ P_j &= p_j + (1-p_i-p_j)p_g/2 \\ P_0 &= 1-P_i-P_j = (1-p_i-p_j)(1-p_g) \end{aligned}\right\} \quad \ldots\ldots (9)$$

And if ties are prohibited (which means that p_g is constrained to unity), we have

$$\begin{aligned} P_i &= (1+p_i-p_j)/2 \\ P_j &= 1-P_i = (1-p_i+p_j)/2 \end{aligned} \quad \ldots\ldots (10)$$

Now, often and perhaps mostly, it is helpful not only to test H_0 ($p_i = p_j$), but to be able to estimate p_i and p_j. In other words the size of the 'indifferent' group, a proportion $1 - p_i - p_j$ of the whole, is a matter of consequence. (If 'i' and 'j' are experimental products a decision on whether either or both should be marketed would turn on such information.) Therefore ties should be allowed in answers. Even so, it does not necessarily follow how estimates of the class sizes should be arrived at. We present the choice of chocolates to a sample of N subjects, and we observe that a number r_i say that they prefer 'i', and a number r_j that they prefer 'j', so that $r_i/N = \hat{P}_i$, and $r_j/N = \hat{P}_j$. But all we have here are two independent equations for three parameters, so that no solution is possible. It is to the credit of G. E. Ferris[8] to have couched this problem in clear statistical terms, to have seen that when replication is introduced a solution *is* possible, and so to have put forward the 'k-visit method'.

The k-visit Method

Let each subject state his preference, in each pair comparison, on k ($k > 1$) different occasions. (Incidentally, if there are extraneous differences, such as varied packs, they can now be balanced out within each subject's k trials.) Any complete (k-fold) result falls into one of 3^k categories, whose expectations can readily be found from the fundamental assumptions and equations. Estimation of the three parameters then becomes possible (although p_g is of no practical interest). For the simplest case, $k = 2$, Ferris gives the maximum-likelihood solutions and their standard errors.

A χ^2 goodness-of-fit test can be applied to the results of the k-visit method, but this must be interpreted conservatively because the model may be spuriously simple. As Ferris himself points out, a more realistic model would assume that many subjects are characterised by a parameter that is the long-run proportion of occasions on which each prefers 'i' or 'j'. A good deal of preference, in other words, is labile. Statistically speaking, the quantities p_i, p_j, and p_g are likely to be the averages of distributions of similar quantities over the population of subjects and occasions. This of course loads the problem with extra parameters and drags it well outside the field of utility. Nevertheless the Ferris model is an intuitively acceptable simplification. Some tests of its *external* reproducibility would be valuable at this stage. It would be good to see, that is, how the spread of replicate estimates, from independent sampling, of p_i and p_j checks with the internal estimate of variance.

Sorting Tests for Item Uniformity

We now return to that common and simple and important question of taste testing for the null hypothesis that two items are sensorily indistinguishable. Pair comparison is merely the simplest of a whole family of experimental designs that can be used for this purpose. The

literature is extensive. A large part of it is devoted to the problem of design efficiency, and not all of this is crystal clear. There is a pervasive belief that because the aim is to test H_0 it follows that relative merits can be assessed on theoretical grounds alone. An assumption frequently made is that if, when the null hypothesis is true, an all-correct solution is statistically more improbable with design X than with design Y, *ergo*, design X is the better. But this situation is rare in practice. Mostly, the null hypothesis is untrue, with, however, only a small difference between the compared items. So it is more realistic to discuss 'betterness' in terms of the latent ability of the design to pick up a small sensory difference.

Consider the general dichotomous sorting model in which we have N_i samples of a product 'i' and N_j of 'j'. They are presented to a subject, in coded random order (or in a coded randomly arranged group), and he is asked to unscramble them by taste differentiation alone. Attention will be restricted to the condition in which the subject is informed beforehand of the size of the N's. We need a probabilistic model that will take into account what happens when the subject is *neither* completely insensitive to the difference between the items *nor* unambiguously recognises them. This situation often obtains: the subject sometimes places an item correctly because his sensory apparatus 'clicks', and sometimes, when there is no 'click', he makes a lucky guess. The process is known in psychology as subliminal perception or *subception*,[7,17] because there is no necessary awareness of which placements are fortuitous and which are subceived. On this basis empirical probabilities of sensory discrimination can be invoked. Let us define p_{si} and p_{sj} as the probabilities of contextual subception of an 'i' or a 'j', respectively. Suppose, as the simplest hypothesis, that $p_{si}=p_{sj}=$, say, p_s. Then it may be shown[11] that the probability of exactly r 'successes' in the sorted-off subgroup of N_i items ($N_i \leqq N_j$) will be a blend of a binomial and a hypergeometric distribution, with probability density:

$$P(r; N_i) = \binom{N_j}{N_i - r} \sum_{x=0}^{r} \frac{\binom{N_i}{x}\binom{N_i - x}{N_i - r}}{\binom{N_i + N_j - x}{N_j}} p_s{}^x (1 - p_s)^{N_i - x} \quad \ldots . (11)$$

An important limiting case is when N_i is unity; then

$$P(r = 1) = 1 - P(r = 0) = \frac{1 + p_s N_j}{1 + N_j} \quad \ldots . (12)$$

When N_j is also unity we have pair comparison, and when it is 2 we have the 'triangle' test. Several higher cases are to be found in the literature.

Power functions can be obtained; i.e., we can seek the probability of the rejection of a false H_0. But the difficulty encountered earlier arises again here: we cannot assume p_s to be independent of design. However, as we said in discussing pair comparison with and without

ties, intuition suggests, and experience substantiates, a negative correlation between the size of p_s and the complexity of the taster's task. With this in mind we may use power functions to arrive at tentative inferences about the merits of rival designs. Note the 'tentative'; here, supremely, the proof of the pudding is in the eating; only experimental observation can supply authentic answers. And we must particularise: findings on ice cream are unlikely to apply to canned soup, and still less to instant coffee. By and large, the safe rule is to aim for simplicity of design.

Discussion and Conclusions

The psychophysicist, preoccupied with intrasubject aspects of the sense of taste, has perhaps traditionally been more inclined to delve into theory than has the worker who is primarily interested in population reactions. Some endeavour is now afoot to unify the two aspects, at least in the matter of model building. No claim to definitiveness is put forward for the psychostatistical models we have discussed. There is patent need for further testing and for comparisons with alternative models.

The main contribution that statistical methodology has made to taste testing is probably in the study of unobvious differences. But whether or not the investigator can establish such a difference, he usually leaves consequent decisions to others. In so doing, he must issue a warning that the detection of a difference is a function of the size, the expertise, and the sensory acuity of the taste panel.

When differences are small, an analytical taste panel operating in splendid isolation—in a socially aseptic environment, as it were—is likely to exhibit greater sensitivity than the consuming public, which does not normally make pair comparisons in any form. We must therefore guard against attaching too much significance to slight differences revealed by a trained panel. A happy corollary is that if a trained panel cannot detect a difference, then neither, *a fortiori*, can the general public. Or is this so? It may be that the consumer population can be regarded as an extremely sensitive tasting system from which all extraneous differences (in the presentation of the items) have been cancelled out. After all, this is in large measure the reasoning behind our making taste panels as big as economy and convenience will allow. Here we are in an area of speculation, and these extraneous factors may not cancel out. Our buying and eating habits destabilise our preferences and probably make us, *en masse*, less sensitive than we are when empanelled for comparative taste testing. Such speculations are admittedly outside the domain of statistical analysis but I shall assume that, for obvious reasons, no apology is necessary for introducing them.

REFERENCES

[1] Bartlett, F. C. (1950). 'Subjective judgements', *Nature*, **166**, 984.

[2] Berkson, J. (1953). 'A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function', *J. Amer. Statist. Assoc.*, **48**, 565.

[3] Bliss, C. I., Greenwood, M. L. and White, E. S. (1956). 'A rankit analysis of paired comparisons for measuring the effect of sprays on flavor', *Biometrics*, **12**, 381.

[4] Bose, R. C. (1956). 'Paired comparison designs for testing concordance between judges', *Biometrika*, **43**, 113.

[5] Bradley, R. A. and Terry, M. E. (1952). 'The rank analysis of incomplete block designs. I: The method of paired comparisons', *Biometrika*, **39**, 324.

[6] Cane, V. R. (1956). 'Some statistical problems in experimental psychology', *J. R. Statist. Soc.*, B, **18**, 177.

[7] Dixon, N. F. (1958). 'Subception', *New Biology* **27**, Penguin Books, Harmondsworth, Middlesex.

[8] Ferris, G. E. (1958). 'The *k*-visit method of consumer testing', *Biometrics*, **14**, 39.

[9] Fisher, R. A. and Yates, F. (1938-57). *Statistical Tables for Biological, Agricultural, and Medical Research.* Oliver and Boyd, London and Edinburgh.

[10] Gridgeman, N. T. (1958). 'Application of quantal-response theory to the cross-comparison of taste-stimuli intensities', *Biometrics*, **14**, 548.

[11] Gridgeman, N. T. (1959). 'Sensory item sorting', *Biometrics*, **15**, 298.

[12] Gridgeman, N. T. (1959). 'Pair comparison, with and without ties', *Biometrics*, **15**, 382.

[13] Gulliksen, H. (1956). 'A least squares solution for paired comparisons with incomplete data', *Psychometrika*, **21**, 125.

[14] Harper, R. (1959). 'Psychophysical methods and laboratory practice', *Laboratory Practice*, **8**, 17.

[15] Jackson, J. E. and Fleckenstein, M. (1957). 'An evaluation of some statistical techniques used in the analysis of paired comparison data', *Biometrics*, **13**, 51.

[16] Kendall, M. G. (1955). 'Further contributions to the theory of paired comparisons', *Biometrics*, **11**, 43.

[17] Lazarus, R. S. and McCleery, R. A. (1951). 'Autonomic discrimination without awareness: A study of subception', *Psychol. Rev.*, **58**, 113.

[18] Maxwell, A. E. (1959). 'Maximum likelihood estimates of item parameters using the logistic function', *Psychometrika*, **24**, 221.

[19] Moir, H. C. (1936). 'Some observations on the appreciation of flavor in foodstuffs', *Chem. and Ind.*, **55**, 145.

[20] Mosteller, F. (1951). 'Remarks on the method of paired comparisons', *Psychometrika*, **16**, 3, 203, and 207.

[21] Scheffé, H. (1952). 'An analysis of variance for paired comparisons', *J. Amer. Statist. Assoc.*, **47**, 381.

[22] Thurstone, L. L. (1927). 'Psychophysical analysis', *Amer. J. Psychol.*, **38**, 368.

THE ECONOMICS OF NEW ROADS WITH PARTICULAR REFERENCE TO M1*

T. K. BURDESS
Ministry of Transport

This article treats the argument for or against building a new road as a problem in economic prediction, and describes methods of obtaining estimates of the data on which predictions can be based.

Introduction

In this country, the money spent on providing new roads or altering existing ones comes from public funds raised either by the Central Government as taxes or by Local Government as rates.

It is interesting to know how this investment compares with other expenditure in the public sector, e.g. on power stations, civil aviation and so on.

This immediately presents a difficulty, since no direct money charge is made for use of the roads as there is for example for consumption of electrical energy, or as fares or freight charges for air transport. A road is not a marketable commodity which is metered for charging the cost to the consumer, as electricity is metered and charged against the consumer in relation to the amount used. The use of aircraft, or indeed any conveyance, can be charged against the user according to the distance the user or his goods are carried.

In other countries, some roads are closed to all but those users who pay a toll charge. It is relatively more straightforward to relate the income from tolls to the various outgoings—the amortised capital investment, maintenance, policing, manning of toll booths, etc.—and derive a sound economic balance sheet.

In this country motorists consider they pay enough for their motoring, and user charges for individual roads would be unwelcome. However it might be noted in passing that there are exceptions. There are the parking meters in Mayfair, which are regarded as putting a justifiable premium upon expensive road space; a charge is made for use of the Mersey Tunnel; and it is proposed to impose a toll charge for use of the Dartford-Purfleet Tunnel, the Severn Bridge, and elsewhere. These tolls will be related to the additional cost of travelling the longer alternative route.

This article is concerned with public roads proposed to be built and used without tolls. In this context, public expenditure on roads is regarded as an investment in the economy of the country, and it is the community of the country which will enjoy the benefits.

* This article is based on a talk given at a Seminar held at the London School of Economics, 18th February 1969.

Benefits

The benefits to the community take several forms:

Journey times may be shortened;
Fuel consumption may be reduced;
Vehicle wear and tear may be reduced;
The number and severity of accidents may be reduced;
New traffic may be encouraged by the new traffic facilities.

With certain reservations and assumptions, it is possible to evaluate these benefits and attribute them to the new road.

It is easier and more accurate to do this after the new road has been built, but by then the need to prove the economic soundness has passed because the funds have already been made available and spent.

In an attempt to measure the benefits before the road is built the following estimates have to be made:

What traffic will be diverted to the new road;
How this traffic will benefit from doing so;
How the traffic not diverted will benefit;
What new traffic is likely to be created and what effect it will have.

A date in the future has to be chosen for this assessment, and it will be a date after the new road is supposed to have been built and opened to traffic.

The value of money has declined year by year and the amount of traffic using the roads has increased. Unfortunately the rate of increase of traffic is neither uniform nor predictable. For example, taking 1954 as 100, traffic has been as follows:

1956	*1957*	*1958*	*1959*
119	123	140	156

Next there is the difficulty of the economic life of the new road. Benefits accruing as soon as the road is opened are worth more than those accruing after the road has been in use for 20 to 25 years, because the latter are discounted at some appropriate rate. For instance, if the rate of interest is 5% the present value of £1 million in 20 years' time is less than £400 000 (£377 000). Nevertheless, the road will still be there and indeed it can be argued that we are still enjoying the benefits of roads built by the Romans! All the same, the further into the future we try to look the more unreliable our estimates become, so, to be able to compare like with like, we have adopted the uniform convention of assessing the benefits about five years ahead. This is about the time it takes to progress from a vaguely planned line on a map to an actual road carrying traffic.

The annual benefits accruing are then related to the total cost of the road. The benefits should exclude taxation, since taxes merely represent transfers of income within the community which provide no real benefit. For example, a saving in petrol should be valued at its price before tax.

I am not an economist so I am not entirely convinced by this argument. Perhaps we should also exclude the taxation element from the cost of building the road—an impossible job of disentanglement! However, it is argued that the revenue derived by the State from taxes levied upon the capital construction would be levied upon the community in any case whether or not this construction work proceeded. Those employed would otherwise be employed elsewhere and would pay tax just the same. On the other hand, the benefits would not otherwise exist nor presumably, would the tax element of those benefits.

Economic Rate of Return

The chief aim is to show the net annual benefits in money terms as a percentage of the capital cost—the Economic Rate of Return.

Economically, a project could be justified if it showed a percentage rate of return not less than could be obtained from other uses of the capital. A reasonable guide might be the interest rate on government stocks, say about 5%. Thus if the annual benefits exceed 5% of the capital cost of the project, the Chancellor of the Exchequer should be interested.

However, our resources, both financial and physical (i.e. construction and materials), are limited and it would not be possible to build at once all the roads likely to show over 5%. They must be graded according to their rates of return and the highest given precedence.

These economic analyses involve the following points.

Traffic

First we need to know about the traffic in the area proposed for the new road. The amount (volume) of traffic can be found by observation on the site or (since this may be laborious and time consuming) there is comprehensive traffic flow information in

The Ministry of Transport National Traffic Census, August, 1954;

The Biennial Metropolitan Police Census in London;

The continuous automatic 50-point census by the Road Research Laboratory;

The monthly manual censuses at the same points taken by the Ministry of Transport;

The annual manual August censuses taken in August by the Ministry of Transport at 100 points.

These latter periodic censuses serve as random sample checks on the traffic trends since 1954.

From these data we can build up a picture of the traffic pattern of the network of roads affected by a new road project. We can also break it down into the following vehicle types:

Cars/Taxis
Buses/Coaches
Light Goods
Medium and Heavy Goods.

But we also need some information about the directional pattern so that we may estimate the changes in pattern that the new road project will cause. This is best done by means of an origin and destination survey in which vehicle drivers are stopped and asked the origin and destination of their journey and sometimes the places of call *en route*.

The new road will reduce the cost of travelling between the centres it serves and some traffic will transfer from the existing roads. This is called diverted traffic.

Some will transfer from one form of transport to another, e.g. train to car. Some will make journeys not previously worthwhile. Transfer from one road to another can be predicted by a traffic assignment procedure, but transfers from rail to road or journeys not previously worthwhile are difficult to predict. This is called generated traffic.

Where the new road is short, generated traffic may be ignored but for a new motorway it may be as much as a quarter in the first year. In bridges and tunnels it is considerably more. For example the Mersey Tunnel opened in 1934 showed during its first year generated traffic amounting to 145% of the estimated diverted traffic. Bridges and tunnels in the United States showed corresponding percentages from 53% to 244%.

Checks are at present being made on M1 to estimate the amount of generated traffic and compare it with predictions.

Speed

To study the reductions in average journey time or the corresponding increases in speed, we need to know:

(*a*) Existing traffic speeds;

(*b*) How speeds will vary with traffic volumes.

We have some novel and interesting ways of finding traffic speeds. For example, we can make timed runs in cars filtered into the traffic stream. This gives us the journey time directly but subject always to the characteristics of the driver and the vehicle. This can be eliminated by making a large number of number-plate checks at each end of the relevant section of road. This method needs a lot of detailed analysis. In small schemes, road junctions etc., this can be overcome by observing traffic from a high vantage point from which vehicles can be traced and timed through a selected length of road, all of which is within view. Then we have speed meters, etc., but these mostly measure speeds at one point.

The Road Research Laboratory scientists have evolved some useful formulae relating speeds to traffic volumes. We are hoping that they will check and enlarge these as they get more new roads to work on. From all these data and with these techniques we should be able to assess:

(1) Vehicle hours per year saved by traffic transferring to the new road;

(2) Vehicle hours per year saved by traffic remaining on affected existing roads because congestion and delay are relieved.
(3) Vehicle hours per year expended by generated traffic; (this will be discussed later).
(4) Vehicle *miles* saved by traffic transferring to the new road. This is sometimes a minus quantity because the new road may be longer than old road as, for example, a by-pass round a town. On the other hand it may save time to travel further and faster on the new road rather than struggle through a congested town centre.

We can then assign unit costs to these quantities according to vehicle types and total the savings in vehicle operating costs.

Accidents

We are entitled to assume that a properly designed new road devoid of frontage building, and all that it entails, is likely to have fewer accidents occurring on it than on the parallel existing roads. We can obtain the accident records for the existing roads affected and we can estimate a comparable accident rate for the new road.

From the accident rates thus obtained we can find the saving in accidents. It would not really be fair to put these results in terms of fatal, serious, slight, and damage only accidents. So we usually give an average cost of £550 per accident saved.

Actually, neglecting court awards and the priceless element of human suffering and bereavement, the economic cost of accidents at 1957 levels were:

Accidents involving personal injury:

Fatal	£2600
Serious	£ 680
Slight	£ 50
Average of above	£ 440
Other accidents (*Damage only*)	£ 50

'Damage-only' accidents are not always reported to the Police but those that are usually number about five or six times the injury accidents.

We therefore usually exclude 'damage-only' accidents and use an average cost of £550 per injury accident to compensate for this omission. To take account of seasonal fluctuations in traffic flow, all the calculations are based upon the average daily flow. Where the average daily flow derives from the Ministry of Transport August census they are multiplied by 300 to give annual values, and those based upon Metropolitan Police censuses are multiplied by 440.

Maintenance

Once a road has been built it creates a new, recurring liability for continual maintenance. The amount spent annually on maintenance is very variable. Even a new road incurs some maintenance costs, e.g.

the Preston By-pass. After teething troubles, however, the cost may be light for some years until the road starts to show signs of wear.

At the risk of erring on the conservative side, we usually make a generous allowance for maintenance. Taking the total cost of maintaining the trunk roads in England and Wales for the latest available year's official return, we divide this by the total mileage (6580) and apply this cost per mile. It works out at about £1500 per mile. This new liability has to be subtracted from the other benefits.

The algebraic sum of all the above amounts can then be expressed as a percentage of the capital cost to give the economic rate of return.

Application of these Considerations to M1

The Road Research Laboratory and Birmingham University combined to assess the economic return of the M1 along these lines. The traffic investigation was carried out in 1955—long before the road was built. There was an Origin and Destination survey covering the traffic on a network of roads between London and Birmingham and from Oxford to the A1 east of Bedford. Traffic was intercepted at 23 points and 41 000 drivers were interviewed. Journey time measurements were made over 1800 miles of road. Assumed speeds on the non-existent motorway were checked against speed measurements on European motorways. It was estimated by this means that (at 1955 levels):

M1 would attract 18 500 journeys a day for an average of 40 miles.

About 30% of all journeys would use 60 miles of M1 but these would contribute just over half of the vehicle mileage on M1.

There would be considerable variations in the traffic flow on M1 with a maximum of 15 000 v.p.d. at Luton and an average of 11 000 v.p.d. (both ways).

About 44% would be heavy goods vehicles and 48% cars.

The A5/A45 between London and Birmingham would be the most important contributing road. Its traffic would be *reduced* by 5000-6000 vehicles per day. The net reduction in vehicle miles per day on existing roads would be 700 000, bearing in mind that there would be an increase on roads giving access to the motorway.

They estimated that 1·6 million vehicle hours per year would be saved by traffic transferring to M1 and the reduced congestion on existing roads would save a further 0·4 million vehicle hours per year. Many of these journeys would *increase* in distance and the total annual increase in vehicle miles would be 13 million.

Accident savings were estimated as 520 casualties per year including 20 fatalities.

It was estimated that the increase in traffic between 1955 and 1960 would be about 30%, and that allowance should be made for generated traffic up to another 30%.

The assessment of economic benefits was done in two parts:

(*a*) benefits to transferred traffic;
(*b*) benefits to residual traffic on existing roads.

The cost of the motorway was taken as £23 300 000 and the maintenance was taken as £200 000 per year. Without the generated traffic the rates of return came out at

3·3% 3·9% 4·2%

according to the three sets of assumed speeds as follows:

Class of Vehicle	First Assignment, m.p.h.	Second Assignment, m.p.h.	Third Assignment, m.p.h.	Speeds on A5/A45 m.p.h.
Private Cars	50	50	50	35
Light Commercial Vehicles up to 1½ tons unladen	40	45	50	31
Medium Commercial Vehicles 1½ to 3 tons unladen	35	40	45	29
Heavy Commercial Vehicles over 3 tons unladen	30	35	40	26

It was assumed that generated traffic (30%) would benefit only half as much as the transferred traffic, so taking account of presumed generated traffic, the rates of return were raised to:

3·8% 4·5% 4·8%

according to the assumed speeds.

These rates are not impressive when compared with a minimum of 5% desired, but they omit any economic value of non-working time saved.

Private car drivers intercepted in the Origin and Destination Survey were asked whether they were travelling as part of their work, for pleasure, or were commuting. By taking in these latter two categories of user, according to the value put on their leisure time, the returns went up to

4·6% 5·4% 7·1%
(leisure valued at 2s. per man hour)
or to 7·8% 9·3% 10·5%
(leisure valued at 10s. per man hour).

Some comparisons of actual with expected figures relating to M1 follow.

Speeds

In the prior assessment speeds of cars assumed were

A5/A45	*Motorway*
35 m.p.h.	50 m.p.h.

Lower speeds given above for various goods vehicles led to the three different rates of return I have quoted.

From an actual sample of 2500 cars in both directions the speeds on M1 were:

Average	*85 percentile*	*95 percentile*
59 m.p.h.	71 m.p.h.	79 m.p.h.

Only two cars in the sample had measured speeds over 100 m.p.h. The average was therefore 9 m.p.h. greater than postulated.

Traffic

The prior assessment concluded that there would be a maximum of 15 000 v.p.d. (at Luton) and an average of 11 000 v.p.d.

In fact, during November there were recorded

17 540 v.p.d. at Friars Wash
and 11 738 v.p.d. at Newport Pagnell.

Cost

As far as we can ascertain, the actual cost (including the Dunchurch By-pass, which was excluded from the prior analysis) amounts to £26 587 484 (nearly £26 600 000).

The above figures are remarkably close to the estimates but I will not try to take comparisons too far, since adjustments should be made to the 1955 figures to bring them up-to-date and also we need a little time for the motorway to settle down.

General Points for Discussion

I have felt for some time that to compute savings in vehicle running costs is a negative approach.

It operates against the stimulus which a new road will give to industrial activity. Someone has said that a road (and a railway too) is an extension of the factory conveyor belt. That being so, the more it is used the better economic proposition it is. This is brought out in the difficulties over estimating generated traffic. Generated traffic is the symptom of the stimulus and can therefore be related to the increase in the National Product which it helps to create.

In other countries, efforts have also been made to assess other indirect benefits such as the rise in value of real estate. In this context we have recently been criticised for building M1 where it is. It connects two prosperous over-developed areas of the country where the production stimulus is not required. Should we not (it is argued) first have built a motorway to South Wales and thus stimulated development of that area of under-employment. I think our answer is that we shall come to that when we have caught up a bit more on the demand which is still frustrated.

REFERENCES

[1] Charlesworth, G. and Paisley, J. L. (1959). 'The economic assessment of returns from road works'. *Proceedings of the Institution of Civil Engineers*, **14**, 229.

[2] Coburn, T. M., Beesley, M. E. and Reynolds, D. J. (1960). *The London-Birmingham Motorway—Traffic and Economics*. Road Research Technical Paper No. 46, HMSO, London.

[3] Roth, G. J. (1959). *Vehicle Operating Costs*. Road Research Laboratory, Road Note No. RN/3426.

[4] Wise, M. J. (1960). 'Some economic trends influencing planning policies', *Journal of the Town Planning Institute*, **46** (2), 30.

[5] Tanner, J. C. (1959). *Accident Rates per Vehicle-mile, Based on August Censuses 1956-1958*. Road Research Laboratory, Road Note No. RN/3608.

[6] Morgan, H. D., Bartlett, V. F., Smith, H. S. and Richards, E. W. W. (1956). 'Toll highways, their economics and construction', *Proceedings of the Institution of Civil Engineers*, **5**, Part III, 575.

[7] Tonti, D. L. (1959). 'N.J. Parkway's economic impact', *Traffic Quarterly*, **13** (2), 198.

[8] Adkins, W. D. (1959). 'Texas Express ways' economic impacts', *Traffic Quarterly*, **13** (3), 333.

INTERVIEWER VARIABILITY AND A BUDGET SURVEY

W. F. F. KEMSLEY
Government Social Survey

This article describes an investigation designed to discover whether the fieldworker introduces a component of variation into the results of a sample survey when the information is obtained from records kept by the informants themselves.

Introduction

In sample surveys using interviews it is well known that, in certain circumstances, the interviewer adds to the variability of the data. It has been demonstrated that this applies to surveys concerned with opinions and attitudes, but is less likely with data on factual subjects.

In surveys dealing with information of an observable and factual nature, the interviewer is unlikely to add to the intrinsic variation of the data. This generalisation applies to questions which are clearly defined and to surveys based on probability samples. If these conditions are not fulfilled, the interviewer may contribute to the variance. A summary of the position will be found in the concluding remarks of a paper by Gray.[1]

These remarks relate to data collected in the course of an interview, and it is only reasonable to inquire whether interviewer variability also appears when the information is obtained from records kept by informants themselves. *A priori*, this may seem unlikely since the interviewer appears to play a subsidiary role. However, if a budget survey using record books is to be properly conducted,[2] fieldworkers must be used to deliver and collect the record books, explain to respondents how to compile and complete their records, and subsequently check through the completed documents. It is conceivable that variations between interviewers in the way they handle the work may influence the final data.

This matter is of great concern to the Social Survey, as the organisation responsible for the fieldwork of the present continuous budget inquiry, known as the Family Expenditure Survey. The study reported below was made in order to find out whether a component of variation was introduced by the field staff, and if so, whether it could be reduced or eliminated by the Social Survey when selecting and training interviewers.

The Material Used for the Study

The current Family Expenditure Survey closely follows the lines of the 1953-54 inquiry of the Ministry of Labour.[3] As in that inquiry,

information is obtained by interview about the composition of the household and the income of each member. Some expenditure is also dealt with by interview, but this is restricted to payments which occur fairly regularly, and the main reliance is placed on record books for details of spending. These are in the form of diaries and cover 14 days instead of the 21 used in 1953-54, but are otherwise similar to the earlier record books.

A multi-stage design is used for the sample, and about 5000 households are covered in a year, divided among 128 primary sampling units. The design is such that although the records of a single household cover only 14 days, there is a continuous flow of records, so that each part of the year is evenly represented. For this study, special arrangements were made for the fieldwork in certain primary units used in 1958 and in the early part of 1959. In each of 24 areas, the pre-selected list of addresses was divided into two equal parts. Two interviewers were used instead of one, but otherwise the fieldwork procedure was undisturbed. The fieldwork of both members of the pair was carried on at the same time and at the same rate, so that the two interviewers had equivalent and inter-penetrating samples. The factors of geography and time would thus affect equally the two half-samples in each primary unit, and in the absence of interviewer variability, differences between the means of the paired interviewers should merely reflect the inherent variability of the data.

A number of variables were examined, and for each the between-interviewer and within-interviewer variances were computed. The two sets of variances were aggregated for the 24 pairs and for four sub-divisions according to interviewer performance, as follows:

TABLE I

Subdivision of interviewer-pairs

	Average Response of Interviewers with: Lower Response %	Higher Response %	Difference	Overall Average Response %
Both had a response above 50% with a difference:				
Group I under 12½%	58	66	8	62
Group II 12½% or more	60	81	21	70
One or both had a response under 50% with a difference:				
Group III under 25%	39	55	16	47
Group IV 25% or more	31	66	35	50

No particular significance should be attached to the precise boundaries of the four groups. They were fixed so as to provide equal numbers of pairs in each group, and at the same time give groups with widely different performances. In particular, groups I and II compare the performance of two interviewers who are both good or fairly good,

whereas group IV contrasts a poor performance with the results from a good response.

The Results

For each variable, the F-ratio, expressing the ratio of the between-variance to the within-variance, was computed. The values are shown in Tables II, III and IV. Table II gives F-ratios for characteristics such as household size and the type of accommodation occupied by the household. These are all factual variables and therefore, on the basis of earlier work, interviewer variability is unlikely to occur. This is supported by the ratios for groups I and II, many of which are below unity. The results for groups III and IV are different; the ratios are higher, and several exceed the 5% level. An interviewer with a poor response obtains data from only some of the households on her half of the pre-selected list. Group IV compares the results from partial co-operation of this kind with the results when all or nearly all cooperate, as was the case with the other interviewer in the pair. The large F-ratios suggest that the results obtained by the interviewer with a poor response differed greatly and sometimes significantly from the interviewer with a good performance. Since the F-ratios for groups I and II confirm that the variable itself is not susceptible to interviewer variability, it follows that the variability arises because the interviewer with a poor response obtains cooperation from only some of her households. It would appear that the group of households from whom she gets cooperation is neither a random nor a representative selection from her part of the list.

The variation imported into the data by incomplete cooperation could arise in various ways. The selection formed by the households which cooperated may be:

(a) systematic and similar as between interviewers;
(b) systematic so far as the individual interviewer is concerned, but differ from one interviewer to another;
(c) incomplete without being systematic.

The first hypothesis postulates a bias common to each interviewer. The difference between the means of a pair of interviewers would then display increased variability when a poor interviewer is contrasted with one having a good response, but not if both have poor responses. With hypothesis (b) or (c) interviewer variability will appear whether both interviewers of a pair are poor, or whether one has a good response; and the F-ratios for group III should then be as large, or larger than, those for group IV.

The results are not consistent with the latter hypothesis; on the contrary, the F-ratios for group III, with one exception, are below those for group IV, suggesting that hypothesis (a) must, to some extent, be true. It seems likely that the selection resulting from a low response is biased in the characteristics covered by the lower part of Table II, and possibly in others outside the scope of the investigation. Incident-

ally, although 'number of rooms' is a factual variable, exact definition is difficult, and this may account for the high ratios in groups II and IV.

TABLE II

F-ratios for various household characteristics

	Both Responses above 50% Difference: Under 12½%	Both Responses above 50% Difference: 12½% or Over	One or both under 50% Difference: Under 25%	One or both under 50% Difference: 25% or Over	All Areas
	Group I	Group II	Group III	Group IV	
Pairs of interviewers	6	6	6	6	24
Within-interviewer variance: degrees of freedom	139	146	94	102	481
Number of persons in household:					
Total	1·12	0·52	0·61	1·96	0·96
Under 16 years	1·71	0·66	0·42	1·72	1·11
16 years and over	0·61	0·37	0·46	1·17	0·65
Working	0·62	0·47	0·28	0·92	0·60
At school	1·00	0·87	0·24	1·90	0·96
All unoccupied	1·77	0·58	0·78	2·59*	1·31
Accommodation:					
Rented/Owner-occupied	0·78	1·47	1·28	4·37*	1·82*
Number of rooms	0·91	4·79*	1·16	6·24*	2·80*
Provision of gas	0·52	0·50	3·43*	1·71	1·17
Other characteristics:					
Occupation of head	0·43	1·34	1·62	3·34*	1·55*
Household bets/does not bet	0·62	0·52	2·29*	2·80*	1·42
Income	0·78	0·69	1·26	1·90	1·13

* Ratio above 5% level

In the upper part of Table II all variables are divisions of the first item. The bias as regards household size would appear to be confined to households containing younger persons and other unoccupied people, and not to extend to older people who are working.

Similar data on expenditure appear in Table III, and from this material it may be concluded that there is no interviewer variability in expenditure recorded by respondents, provided the response is good. On the other hand, the secondary effect produced by comparing in group IV a poor interviewer with a good one, also appears in the *F*-ratios for total expenditure (first line of Table III).

The values in the first line are consistent with the hypothesis that a poor response leads to bias in the composition of the achieved sample, and a consequent distortion of the expenditure data. One characteristic which is biased by a poor response is the income distribution of households, with which of course expenditure is correlated. Moreover, some parts of the expenditure field are directly associated with the variables in Table II; in particular, housing costs and expenditure on gambling, and these will automatically contribute to interviewer

variability. Excluding these items, the remaining expenditure has been subdivided into 13 groups, and the corresponding *F*-ratios are shown in Table III. Only two categories, namely the running costs of motor vehicles and travel by public transport, show large ratios for group IV. Possibly this is the result of another bias in the sample, namely whether the household uses a private car or public transport, but direct evidence on this point is not available. Apart from these two categories, it is difficult to reach a clear-cut conclusion. However, whereas in groups I and III there are approximately equal numbers of *F*-ratios under 1 and over 1, and group II has substantially fewer above 1 than below 1, there are in group IV only three ratios below 1, but 10 above. This suggests that the effect of biases in the composition of the sample brought about by a low response rate may not be confined to a few categories of expenditure, but may extend to other sectors.

TABLE III

F-ratios for expenditure recorded by the whole household

	Both Responses above 50% Difference: Under 12½%	Both Responses above 50% Difference: 12½% or Over	One or both under 50% Difference: Under 25%	One or both under 50% Difference: 25% or Over	All Areas
	Group I	Group II	Group III	Group IV	
Total expenditure	0·95	0·27	1·03	2·35*	1·10
Expenditure on:					
House maintenance	1·17	1·66	1·11	1·27	1·28
Fuel and light	0·77	1·65	0·42	1·18	1·00
Food	1·42	0·30	0·78	1·30	0·93
Alcoholic drink	0·13	0·75	1·38	0·34	0·63
Tobacco	0·79	1·35	1·45	0·32	0·92
Clothing	1·56	1·83	1·23	1·40	1·51
Footwear	1·27	0·92	0·64	0·95	0·94
New durables	0·78	0·72	1·10	1·61	1·07
Vehicle running costs	1·62	0·35	0·96	2·15	1·15
Public transport	1·45	0·55	0·55	2·43*	1·27
Recreational, chemists' and personal goods	1·49	0·70	0·95	1·40	1·15
Insurance	0·39	1·21	0·82	1·28	0·93
Other, excluding housing costs, purchases of motor vehicles or houses, and betting	0·67	0·55	1·56	1·81	1·15

* Ratio above 5% level

This technique can also be applied to throw some light on certain technological problems. Two instances will be quoted. In most, if not in all budget surveys, expenditure exceeds income for the sample as a whole, and also for various breakdowns of the sample. In other words, the ratio of expenditure to income will usually exceed 100%. The first line of Table IV gives the between-within *F*-values based on this ratio and shows that for the Family Expenditure Survey the variation in this ratio does not contain an element due to interviewer variability. The second instance concerns the variation in expenditure between the

separate weeks of the record. As I mentioned in an earlier paper,[2] the level of recording tends to fall after the first three days, with the result that expenditure in the second and subsequent weeks is below that reported for the first week. The last two lines of Table IV have a bearing on this inter-week variation. The penultimate line was derived by calculating the between-within F-values based on the ratio of the number of entries made in the first week to the number in the second week. The variation in the ratio between interviewers is no more than would be expected by chance from the variance within interviewers. Similarly, the ratio of the number of entries in the first three days of the first week to the corresponding number in the second week was studied. This, however, shows some degree of interviewer variability.

TABLE IV

F-ratios for some variables of technological interest

	Both Responses above 50% Difference: Under 12½%	Both Responses above 50% Difference: 12½% or Over	One or both under 50% Difference: Under 25%	One or both under 50% Difference: 25% or Over	All Areas
	Group I	Group II	Group III	Group IV	
Ratio of expenditure to income	0·90	0·90	0·75	0·21	0·70
Number of entries					
Total recorded by household	1·10	0·17	0·47	0·63	0·58
Ratio of first week to second week:					
Whole week	1·48	0·31	0·48	0·55	0·65
First three days of each	1·88	0·91	0·22	2·06	1·12

In the two technological problems just discussed it seems fairly safe to conclude that any internal bias that may exist in an individual record cannot be attributed, in general, to the interviewer, except possibly for variations in the first three days' entries.

Conclusions

The main finding is that there is no evidence of interviewer variability in the data on expenditure recorded in a budget survey, provided that a good response is achieved by each interviewer individually; say over 60%. This conclusion, however, is based on a survey in which experienced Social Survey fieldworkers were employed. Whether this would hold if inexperienced personnel were employed is not known.

On the other hand, if the response is poor, interviewer variability appears when an interviewer giving a poor performance is compared with one providing a good response. It seems that the sample achieved by someone with a poor response is not a representative selection of the pre-selected list on which the interviewer is working, but is biased in a number of directions. While there is no reason to believe that this bias is caused by any deliberate or conscious act of the interviewer,

nevertheless the effect is rather as though a biased sample had been used. A low response will tend, therefore, to make ineffective much of the work put into designing and selecting a probability sample. In some respects the situation would then resemble that obtaining with quota samples.[4]

The importance of a satisfactory sample is, nowadays, generally accepted, and from the preceding discussion it would appear that to achieve this, each interviewer must secure a satisfactory response. Therefore, in selecting interviewers for this inquiry the Social Survey endeavours to use only those who can be relied on to produce a good response, as well as deal with the interviewing side of the survey. By careful selection it is possible, as we have confirmed by experience, to improve the overall response considerably. However, it is not an easy matter to make this selection; so far, the only reliable guide we have found is actual performance in the field over a period of time. It would seem that this is a problem of selection which would repay further research.

REFERENCES

[1] Gray, P. G. (1956). 'Examples of interviewer variability taken from two sample surveys', *Applied Statistics*, **5**, 73.

[2] Kemsley, W. F. F. (1959). 'Designing a budget survey', *Applied Statistics*, **8**, 114.

[3] Ministry of Labour and National Service (1957). *Report of an Enquiry into Household Expenditure in 1953-54*. HMSO, London.

[4] Moser, C. A. and Stuart, A. (1953). 'An experimental study of quota sampling', *J. R. Statist. Soc.*, A. **116**, 349.

NOTES AND COMMENTS

We regret that there is insufficient space in this issue to allow the usual list of meetings of Sections and Local Groups to be brought up to date. However, readers may be interested to know that another one-day conference was held on May 18th by the Birmingham and District Group at the College of Advanced Technology, Birmingham. The subject on this occasion was Dynamic Programming and individual papers were contributed by Dr M. G. Simpson (An Introduction to Dynamic Programming), Dr S. Vajda (The Theory and Practice of Dynamic Programming), Dr G. Morton (Comparing Dynamic Programming with other Methods), and Mr D. V. Lindley (Dynamic Programming and Decision Theory). The conference was very well attended; it seems that this form of meeting is well worth the time and trouble taken in organising it.

* * *

The Institute of Management Sciences has sent details of the Seventh International Meeting to be held in the Hotel Roosevelt, New York City on 20th-22nd October 1960. The general topics to be discussed are: computers and simulation techniques; management science—human aspects, techniques, mathematical methods and tools, significant application; behavioural science and management science. Further information may be obtained from Mr James Townsend, Union Carbide Corporation, 270 Park Avenue, New York 17.

Also from the Institute of Management Sciences comes the first issue of *Management Technology*, a monograph on case histories; it is apparently intended to publish further monographs from time to time. The monograph includes: (1) case histories of the applications of management sciences to management problems ;(2) papers on the problems of applying new developments in management sciences to business systems; and (3) management discussions of management sciences, such as the nature of business problems which they feel are important and can be attacked by management sciences. The editors are Roger R. Crane and C. West Churchman.

BOOK REVIEWS

Forman S. Acton: *Analysis of Straight-Line Data*. New York, John Wiley and Sons; London, Chapman and Hall, 1959. Pp. xiii + 267. 72s.

E. J. Williams: *Regression Analysis*. New York, John Wiley and Sons; London, Chapman and Hall, 1959. Pp. ix + 214. 60s.

These two books cover similar topics and there is a certain amount of overlapping, though not as much as might have been expected. Dr Acton is a professor of electrical engineering and a mathematician; Dr Williams is a statistician and was formerly a professor at the University of North Carolina; both claim to be writing for the experimental scientist with some knowledge of statistics rather than for the statistician.

I liked the start of Acton's book and thought that it showed a realistic approach, but I fear that the loss of enchantment with the later material, experienced by the author, afflicted me at an earlier stage. Chapter 2, dealing with classical regression analysis with x free from error, struck me as somewhat laboured and cluttered up with inessentials. I think a chapter of this length defeats its own object, and encourages a scientist to claim that he has managed without these techniques in the past and does not see that it is necessary to trouble himself with all these complications. Actually the really important techniques can be, and have been elsewhere, explained in quite a short space without descending to the recipe book level; the more recondite techniques could have been put in an appendix and the approximate methods transferred to another chapter. Chapter 3 deals with the case where there are several y values for each x and Chapter 4 with samples from bivariate normal populations; no particular comment is called for here.

Chapter 5 moves on to the case whether both x and y are in error, but I find it disappointing. The chapter heading is misleading: it speaks of regression whereas the text deals with the estimation of the functional relationship. In this connection one might have expected some reference to the long paper by Lindley which concludes that the important class of problems in which y has to be predicted from an observed value of x can be satisfactorily tackled by calculating the regression of y on x, the latter assumed free from error. Of course, Acton may dispute this, but it is a commonly accepted method and ought at least to have been discussed. In his treatment of the functional relationship Acton assumes that the errors in x and y may be correlated and spends most of the chapter dealing with the case where the data occur in groups and the degree of error correlation is known. He does this on the grounds that it is rarely safe to assume the errors to be uncorrelated and that in any case this is merely a special case of the formula he deduces. This approach seems to me to be unduly rigorous; I think it would have been helpful to quote Lindley's equation applicable to uncorrelated errors, since it is simple to use, and does not assume grouped data. It is arguable that instances where Lindley's equation is a good approximation may be more numerous than those in which the extra knowledge presupposed by Acton is available. Moreover, calculations on one of

the examples show that the difference between Lindley's line and Acton's (assuming error correlation about 0·7) is small compared with the confidence interval. I suspect that this would often be the case for reasonably precise physical data, where the gradient of the functional relationship is most often required.

In equation (10) on page 135, the second bracket should contain R^2 instead of R, and the numerator of the right-hand side of the equation above should be multiplied by βR. The treatment of confidence limits confused me; three methods were used, leading to different results, and I think that any reader may be pardoned for feeling that this section lacks a real sense of conviction. In fact I believe, though this is not the place to give the details, that a more satisfactory method is probably available for calculation of the line and its confidence limits. But, to return to a previous point, there must be many cases where we are not interested in the functional relationship, but only in predicting y from measured x. Here intuition would lead me to use the normal regression of y on x, as advocated by Lindley. This deals with the measured value of x as the object of study and therefore as without error; hence the correlation of the errors of x and y (which refer to deviations from true values) is irrelevant. Also it would be appropriate to use the Berksonian technique here; why was it mentioned in Chapter 2 rather than in this chapter?

Later chapters deal with the analysis of variance with several lines, orthogonal polynomials, the use of transformations, the rejection of unwanted data, and cumulative data, though the last three chapters are very brief. If the amount of detail given in this review to the various parts of the book seems rather out of proportion, it may be put down to the fact that much of the book covers fairly familiar ground, whereas Chapter 5 leaves the beaten track and is more provocative.

After an introductory chapter concerning such things as significance tests, Dr Williams deals briefly with linear regression and with a simple type of weighted linear regression, though this chapter does not contain all he has to say about various related problems. He then passes straight on to multiple and polynomial regression, orthogonal polynomials, and equations with coefficients subject to linear restrictions. At this point I may venture the opinion that the exposition, though reasonably clear, might sometimes be made more so; in particular, more 'illustrative working-out' for Crout's method of matrix inversion in the example where it is first described would make all the difference between slight uncertainty and crystal clarity.

The following chapter describes regression equations requiring iterative calculation, such as exponential and hyperbolic regressions, and linear regression with estimated weights. Chapter 5 discusses the problem of choosing between two or more possible regression functions. Some of the topics omitted from Chapter 2 are discussed in Chapter 6, which deals with estimation from the regression equation. Williams uses a fiducial argument here and elsewhere in the book, and considers not only simple estimation but estimation of maxima and minima. After a short chapter on covariance analysis, there is a more substantial one on the treatment of heterogeneous data. This includes some interesting topics: tests of differences in gradient and position of regression lines; tests of concurrence of lines; tests of proportionality of regression lines; the comparison of regression equations from sets of correlated data, etc. Some of these topics are rarely, if ever, discussed

in textbooks. Chapters 9 and 10 are concerned with simultaneous regression equations and discriminant functions respectively. The last chapter considers functional relations; although these are outside the main field covered by the book, they are of some importance in scientific work and are conveniently dealt with in a book of this kind. Lindley's 1947 paper is referred to, but Williams treats the problem of the estimation of the functional relationship in an apparently different way. There is nothing wrong in this, of course, though it seems to me that Lindley's method is more convenient to use. But it seems a pity not to have stated that the method, though superficially different, is algebraically equivalent to Lindley's, and that the determination of fiducial limits is algebraically equivalent to a method published by Miss Creasy in 1956.

A comparison of these two books is hardly possible, because they set out to do different things, but nevertheless it is convenient to consider them together. Acton gives a good deal of information concerning linear regression, but the reader needs to have a certain amount of patience to sift the really useful material from that of incidental interest. For a book of its length and subject, I thought it had too few down-to-earth practical examples. Williams was better in this respect; there were many more examples and the book impressed me as having come to grips with quite a variety of interesting experimental situations. This variety results in a fairly concentrated expository style; at times it is not easy reading. As the author states in the preface, some familiarity with statistical methods is assumed. I should be inclined to regard the simpler sections as a textbook and, after a preliminary skimming, to regard the more difficult sections as a reference book to be pondered over when need arises.

D. G. BEECH

PUBLICATION RECEIVED

The Public and the Programmes. London: British Broadcasting Corporation, 1959. Pp. 71. 8s. 6d.

This Report is based on an inquiry carried out at the end of 1958 by the BBC's Audience Research Department. The inquiry was a special project, designed to supplement the information regularly made available to the Corporation as a result of the day-to-day operations of the department. It is now published in the belief that it contains information of general interest. The first chapter deals with such matters as listening and viewing facilities, the composition of the sound and TV public, and the proportion of people who see the *Radio Times* and those who see a commercial TV journal. The following chapters deal with listening in the evening, viewing in the evening, listening in the daytime, and the tastes of listeners and viewers. Then follow a summing-up entitled 'Highlights and Reflections', and a brief Appendix on the sampling plan and fieldwork results. The report is presented with a wealth of tables and diagrams, and contains much material of interest to sociologists.

STOCK CONTROL PROBLEMS: SOME FALLACIES IN THEIR CURRENT TREATMENT

ERIC DUCKWORTH

In this article Mr Duckworth has not attempted a comprehensive review of stock control literature. Instead he shows how easy it is for fallacies to arise in the mathematical treatment of stock control problems, with results that might be extremely costly if they were applied to stock management.

Stock Control problems are among the major ones to be tackled by industrial statisticians. Their importance can be judged by the size of the stocks held by industry. One estimate[1] gives a figure of £4000 million and suggests that this amount could safely be reduced by £1000 million if the problems were tackled vigorously and scientifically.

The statistical literature on this subject is quite comprehensive and yet the number of reported cases in which statistical techniques have influenced the practical management of stocks appears to be few. In my view this may be because the current statistical and mathematical treatments contain fallacies as a result of which their application can produce an excessive level of stocks. It is the purpose of this paper to reveal these fallacies and suggest a more correct treatment.

The two major problems in any stock control system are (*a*) to decide what should be the minimum stock level of any commodity at which a fresh order for that commodity should be placed; and (*b*) to decide upon the amount to order each time replenishment is necessary. The system considered in this paper will be a constant re-order quantity system because it has been shown that this generates less average stock than a constant re-order interval system.[2] The same fallacies occur in both systems, however, and so it is only necessary for the moral to be pointed in one case.

Minimum Stock Level

The usual system postulated is that shown in Fig. 1. Ideally, replenishment of stock occurs at points A, B, C, D and the amount of

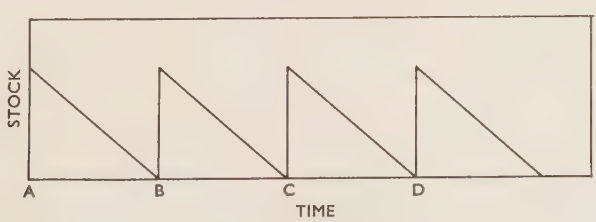


Fig. 1. Stock control system; simple deterministic model with no buffer stock.

stock received is just sufficient to cover the sales in the periods A-B, B-C, C-D, etc.

Sales are not uniform, however, and there is uncertainty about when the stock will decline to zero and replenishment will be necessary. Replenishment is not instantaneous and orders have to be placed some time in advance of the expected moment of zero stock. The actual situation is thus more realistically described in Fig. 2. (I am indebted to Mr A. Battersby for this particular diagram.)

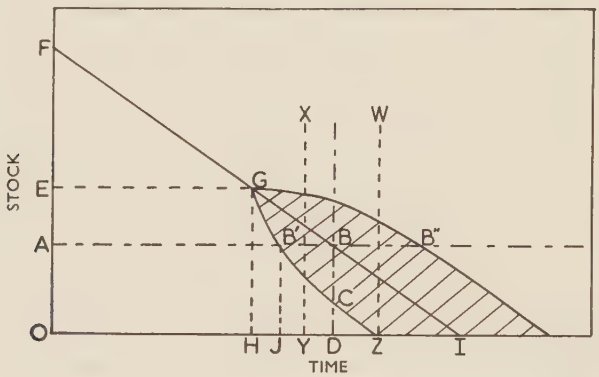


Fig. 2. Stock control system; more realistic model with buffer stock.

In the diagram the points A and B are the same as in Fig. 1. The expected decline of stock with time is shown by FB and if stock could be replenished instantaneously then, as before, no orders need be placed until stocks are exhausted at point B. (I have, however, come across cases in the grocery trade where replenishment from warehouses is virtually instantaneous and yet quite high buffer stocks are carried in shops. This suggests that a fruitful field for work on stock control exists in retail business.)

Stock cannot, in general, be replenished instantaneously and must be ordered at, say, point G (time H) if fresh stocks are to be received at point B (time D). In the time HD during which replacement stock is on order there is uncertainty about the actual sales, and thus about the decline in stocks, which will occur. This uncertainty is shown by the shaded area enclosed by the lines GB′ and GB″. There is, let us say, a 95% probability that the decline in stocks during time HD will be within the shaded region. There is therefore a $2\frac{1}{2}\%$ chance that the stock EA will decline to zero as early as or earlier than time J, i.e. at the point B′. To cater for the eventuality of this stock running out between J and D, a level of buffer stock equal to BC is carried, so that the new stock EC has a high probability of lasting until time D at which replacements arrive.

It has been known for statistical investigation to stop at this point and for the buffer stock level to be calculated purely on the basis of the variation in sales rate. This is the first fallacy which must be pointed out, because there is another kind of variation to be taken into

account, the variation in the delivery time HD. Let us say there is a 95% probability that delivery of replacement stock will be made within the time YZ, as indicated by the dotted lines XY and WZ, and that time D is the median of the distribution. The chance that stocks will be exhausted before replenishment is thus more than $2\frac{1}{2}\%$ even when a buffer stock of BC is carried. To reduce the chance of a stock-out to $2\frac{1}{2}\%$ again, a buffer stock of size between BC and BD will have to be provided in order to allow for the possible combined eventuality of a decline of stocks at least as fast as the line GBCZ and of replenishment no earlier than time Z.

To reduce this combined probability below $2\frac{1}{2}\%$ even larger buffer stocks will have to be maintained. In designing a stock control scheme based on the considerations of Fig. 2, the statistician should also beware that the combined possibility that the stock may decline more slowly than line GB″ and a delivery may be received earlier than time Y does not result in the accumulation of more stock than the store will hold.

When the combined probability of a stock-out has been calculated for a given situation, this is usually interpreted by the statistician as the risk that the particular commodity will be out of stock at any moment of time. Where a store contains many commodities which are varieties of a certain class of parts, e.g. bearings, or nuts and bolts, whose rates of sale (or withdrawal from store which is fundamentally the same thing) can be shown to follow the same distribution, and whose delivery dates can similarly be demonstrated to form one distribution, then this combined probability is taken as the *proportion* of those commodities which will be out of stock at a given time.

This is the second fallacy, because the combined probability is relevant only to commodities within the shaded regions of Fig. 2. That is, it will apply only to those items (where an item means a collection of a particular commodity, say a particular size of bolt) whose stocks have reached the point G at which a fresh order for replacement must be made. The height AF is related to the size of replacement batch which it is most economic to order (this will be discussed later), and where this is large compared to the coefficient of variation of sales rate and of delivery time the number of items for which replacements are on order will be small; although the risk that one of these items could be out of stock before replacement arrives may be $2\frac{1}{2}\%$, the proportion of the total number of items found out of stock at any time will be much less than this.

Contrariwise if the economic replacement batch is small, possibly because of transport capacity restrictions as in the case, say, of oil tankers replacing refinery stocks, then an individual replacement batch may be very much less than AE in height and thus the proportion of items likely to be out of stock very much higher than that given by the combined probability. Only where the replacement batch equals AE in height will the above calculated probability be the same as the proportion of items expected to be out of stock.

A stock manager is more concerned with the proportion of his total stock items unavailable at any given time than with the probability that the items awaiting replenishment are out of stock. Thus the latter probability is not usually very helpful to him. The difference between this probability and the proportion of items out of stock will now be illustrated mathematically.

Proportion of Ordered Items out of Stock Before Replenishment

The risk of running out of stock of any item on order can be obtained by calculating the frequency distribution of the amount of stock at the expected point of replenishment, B in Fig. 2, and then dividing by the expected rate of sales to obtain the distribution of the life of the stock. The area of this distribution curve which is below zero life is the required probability, if we adopt the usual convention that the total area under the curve is unity. This stock life distribution can be obtained for particular items directly from previous records of stocks on receipt of delivery, if these are sufficiently numerous, or by a combined sampling from the distributions of sales and of delivery times.

Where there are only a few items in each store this calculation can be carried out for each item. Where there are 2000 items or so, a situation not unusual in industry, certain simplifying assumptions usually have to be made.

The first assumption is the one mentioned above in connection with the bearings and nuts and bolts, that the sales of the various items are assumed to belong to the same distribution. This in many cases is clearly not true; large bolts are consistently used in smaller quantities than small bolts. If, however, the actual sales are divided by the expected sales—and values of expected sales for every item must be obtained in order to run a stock control system—then it is often found that this new variable can be assumed to have the same distribution for each item. It is often a skew distribution because by definition it cannot have any values less than zero, and a lognormal distribution usually fits satisfactorily.

The second assumption is that delivery times for each item can be described by one distribution. This is very often true. The following mathematical relationships can then be derived.

Let q = stock remaining when fresh delivery is required
l = factory delivery time (HD in Fig. 2)
$\bar{l}$ = average factory delivery time
σ_l = standard deviation of l
t = buffer stock (BD in Fig. 2) in terms of expected sales rate
s = actual sales per unit time
$\bar{s}$ = expected sales per unit time
σ_s = standard deviation of $\bar{s}/s$

Then q = expected interval between order time and delivery time multiplied by expected sales per unit time, *minus* actual interval between order time and delivery time multiplied by actual sales per unit time, *plus* buffer stock in terms of expected sales rate multiplied by expected sales per unit time.

$$q = \bar{l}\bar{s} - ls + t\bar{s}$$

$$\frac{q}{\bar{s}} = \frac{(\bar{l}+t)\bar{s} - ls}{\bar{s}} \qquad \ldots . (1)$$

Let $W = (s - \bar{s})/\bar{s}$

Then $s = \bar{s}(1 + W)$

Substituting for s in equation (1) gives

$$\frac{q}{\bar{s}} = \frac{(\bar{l}+t)\bar{s} - l(1+W)\bar{s}}{\bar{s}}$$

$$= \bar{l} + t - l(1+W)$$

This distribution of $q/\bar{s}$ is very difficult to evaluate analytically because it involves the product of two parameters l and W which are themselves distributed. I am indebted to Mr D. Farlie for the suggestion that the problem could be tackled by considering the logarithm of the product term. The characteristic function for the logarithm of the product term would be the product of the characteristic functions for the logarithms of l and $(1 + W)$ and the desired distribution can be obtained by solving the inversion integral. Alternatively the characteristic function can be used to fit the best Pearson type curve by the method of moments and then the exponential of the variable in the Pearson curve can be extracted to give an approximating function for the desired distribution. In either case the mathematics is somewhat intractable and the resulting function likely to look very unwieldy.

In most cases known to me the solution is usually obtained by sampling from the two distributions and computing numerically the distribution of the product of the sample values.

An ingenious approach by S. Babik simplifies the mathematics, however. Instead of considering the distribution of stock life in terms of expected sales, $q/\bar{s}$, let us consider the stock life in terms of actual sales, q/s. There is no reason why the one should not be as good as the other for our purpose.

Then $q = \bar{l}\bar{s} - ls + t\bar{s}$, as before

and

$$\frac{q}{s} = \frac{(\bar{l}+t)\bar{s} - ls}{s} \qquad \ldots . (2)$$

Let $(\bar{s} - s)/s = x$

Then $\bar{s} = s(1 + x)$

Substituting for $\bar{s}$ in equation (2) gives

$$\begin{aligned}\frac{q}{s} &= \frac{(\bar{l}+t)s(1+x)-ls}{s} \\ &= (\bar{l}+t)(1+x)-l \\ &= (\bar{l}+t)x+\bar{l}+t-l \\ &= (\bar{l}+t)x+t-(l-\bar{l})\end{aligned}$$

The distribution we seek is thus the sum of two known distributions, and the amount of mathematics is greatly reduced.

In one case studied by me x was normally distributed with zero mean and $(l-\bar{l})$ was also normally distributed with zero mean.

Thus q/s was normally distributed about t with

$$\text{variance} = (\bar{l}+t)^2\sigma_s^2+\sigma_l^2$$

This situation probably occurred because the variance of s was small compared to $\bar{s}$ and so was the variance of l compared to $\bar{l}$. This situation may be by no means unusual in industry however, and the distribution of q/s may be normal or nearly normal in many other cases.

The distribution is illustrated in Fig. 3.

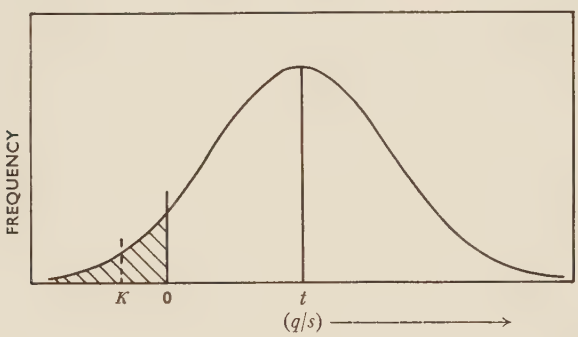


Fig. 3. Assumed distribution of stock life in terms of actual sales, q/s.

The area shaded to the left of $q/s=0$ gives the required probability of an item running out of stock before the new delivery is received. It can be obtained directly from tables of the areas under the normal curve, e.g. Table IX in 'Fisher and Yates', once the values of $\bar{l}$, t, σ and σ_l have been determined.

In the above case $(\bar{s}-s)/s$ was distributed normally about zero with standard deviation $\sigma_s=0{\cdot}35$.

The factory delivery times were distributed about the expected

TABLE I

Probability of running out of stock for an item on order

Expected Factory Delivery Period in Months	Buffer Stock in Months of Expected Sales: 1	2	3
3	26·4%	14·7%	8·9%
4	30·0%	18·5%	12·1%

value with standard deviation $\sigma_l = 0{\cdot}75$ months. By choosing values of $t = 1, 2$, and 3 months and of $\bar{l} = 3$ and 4 months Table I was constructed.

All these probabilities seem very high and would probably be considered too high by any stock controller who regarded these probabilities as representing the proportion of his stock likely to be unavailable at a given time.

As stated earlier they do not mean this, although they are often represented in the literature as so doing. They represent only the probability that an item on order from the factory will be out of stock before the order is fulfilled. The probability that an individual item, whether on order or not, will be out of stock at a given time is much less than in this Table, as will now be shown.

Proportion of All Items out of Stock

Let the orders from the factory be delivered at discrete intervals of time (e.g. weekly) at the constant rate of $\mathcal{N}$ per unit of time (the following argument is not invalidated by a distribution of $\mathcal{N}$ and is simplest to follow by assuming a constant $\mathcal{N}$).

Also let the area to the left of any time K in Fig. 3 be p_K. In this notation the probabilities listed in Table I above are p_0.

At time 0, just before a factory delivery, let there be I items out of stock. When the delivery is received at time 0, $p_0\mathcal{N}$ items are filled in, because p_0 is the proportion of the items on order which were out of stock just before delivery. The remainder of $I - p_0\mathcal{N}$ items are left to be re-stocked on subsequent deliveries.

The items delivered in period 1, which were also out of stock in period 0, will have been out of stock for 1 or more periods. There are $p_1\mathcal{N}$ such items. Similarly for periods 2, 3, ... until all items I have been filled in.

Hence $I = \mathcal{N}(p_0 + p_1 + p_2 + \ldots . + p_\infty)$

This relationship $I = \mathcal{N}\Sigma p$ assumes that the number of items filled in during week i will be $p_i\mathcal{N}$. This is strictly true only if the items which arrive during any one week are a random selection from all items outstanding. This assumption is reasonably justified when the standard deviation of factory delivery time is fairly small. Where the standard deviation is high, then priorities may have to be applied to urge the more outstanding items and the assumption may no longer be valid. For this paper, which is mainly concerned with illustrating principles, the assumption will be used.

The relationship $I = \mathcal{N}\Sigma p$ also seems to produce the apparent paradox that the greater the number of items ordered per week the greater the number of items out of stock. This is not really a paradox because $\mathcal{N}$ is not an independent variable; it depends, as shown below, on the total number of items in the store and the amount of stock of each held.

Let total number of items $= T$.
At the steady state, when N items are ordered per week, stock of each item, in terms of sales (i.e. stock divided by expected sales) excluding buffer stock, $= T/N$.

$$\text{Let } T/N = H$$
$$\text{Then } N = T/H$$
$$\text{and } I = (T/H)\Sigma p$$

The apparent paradox is then resolved because for a given Σp and H, the number of items out of stock increases with the total number of items in the store. For a given Σp and T, I increases the smaller the stock, H, of each item. This clearly illustrates the point made earlier that the chance of an individual item being out of stock depends not only on the buffer stock, but also on the total stock held, because this determines N, the number of items on order at a given time.

TABLE II

Calculation of Σp

$-z$	p_0 to p_{15}
0·630	0·264
0·787	0·218
0·944	0·173
1·101	0·135
1·258	0·104
1·415	0·079
1·572	0·058
1·729	0·042
1·886	0·030
2·043	0·021
2·200	0·014
2·357	0·009
2·514	0·006
2·671	0·004
2·828	0·002
2·985	0·001
	$\Sigma p = 1{\cdot}160$

If the previous example is used again, in which $N = 60$ items per week, and the time interval is taken weekly in the calculation of $N\Sigma p$, then $p_0 \ldots . p_\infty$ can be calculated as below.

Take $\bar{l} = 3$, $t = 1$, $\sigma_s = 0{\cdot}35$, $\sigma_l = 0{\cdot}75$, all in months.

Then $$\sigma_y = \sqrt{\sigma_l^2 + (t+\bar{l})^2\sigma_s^2} = 1{\cdot}588 \text{ months}$$
$$= 6{\cdot}352 \text{ weeks}$$

At time 0, $-z$, the distance from the mean in standardised units, is t/σ_y.

$$t = 4 \text{ weeks}$$
$$\text{Thus } -z = \frac{4}{6{\cdot}352} = 0{\cdot}630$$

The area to the left of $-z = 0{\cdot}630$ on the normal curve is 0·264.

This is p_0. One week later $-z = \frac{5}{6 \cdot 352} = 0 \cdot 787$.

The area to the left of $-z = 0 \cdot 787$ on the normal curve is 0·218. This is p_1.

In similar fashion p_2, p_3, p_4, etc. are calculated by taking 6, 7, 8, etc. as the numerator in the expression for $-z$; the values of $-z$ and p are collected in Table II.

With 60 items per week $\mathcal{N} \times 1 \cdot 160 = 69 \cdot 6 \simeq 70$.

By using similar calculations for $t = 2$ and 3 months and $\bar{l} = 3$ and 4 months the following table is built up:

TABLE III

Number of items out of stock

Expected Factory Delivery Period in Months	Buffer Stock in Months of Sales: 1	2	3
3	70	39	25
4	96	60	40

With 2000 items in the entire store the percentage of items unavailable at any time is as follows:

TABLE IV

Percentage of items unavailable at any time

Expected Factory Delivery Period in Months	Buffer Stock in Months of Sales: 1	2	3
3	3·5	2·0	1·3
4	4·8	3·0	2·0

These are far smaller than the ones given in Table I and are much more workable from the point of view of a stock controller. In this actual example $\bar{l}$ was arranged to be 4 months and a buffer stock of 2 months was provided. The scheme appeared to work satisfactorily until circumstances changed and a further study was needed.

Simulation Studies

To check on the result of the above calculations a simulation study was carried out. The conditions chosen were $t = 2$ and 3 months and $\bar{l}$ was arranged to be 4 months as above.

For the purpose of the simulation it was assumed that all orders for a week were placed at the beginning of the week. The actual number of orders placed per week was 6, and each item ordered had a stock of $\bar{l}\bar{s} + t\bar{s}$ at the time of ordering. The number of items out of stock found from the simulation was then multiplied by 10 to accord with the theoretically analysed situation of 60 orders per week. The rate of sales for each item over the delivery period was sampled from a normal distribution with the mean of $x = (\bar{s} - s)/s = 0$ and $\sigma_s = 0 \cdot 35$. The delivery period was given a standard deviation of 0·75.

When an item ran out of stock the length of time it was out of stock before replenishment was marked on the type of chart shown in Fig. 4.

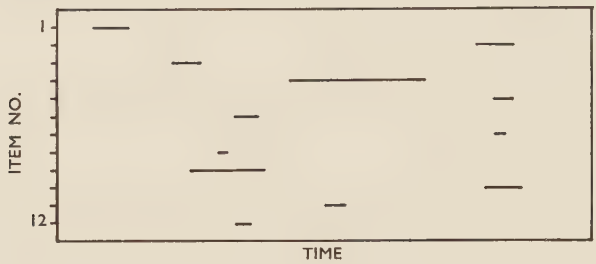


FIG. 4. Type of chart used in simulation studies.

From this chart, run over a sufficient length of time, it was possible to compute the probability of a stock-out and the average number of items out of stock. This was done and the comparison with the theoretical result is shown in Table V.

TABLE V

Comparison of simulation studies with theoretical results

	t	Theoretical Result	Simulation Result
Probability of Stock-out	2 3	18·5% 12·1%	19·4% 14·2%
No. of Items out of Stock	2 3	60 40	68 38

The simulation result agreed well with the theoretical. Other simulations have been carried out which give equally good agreement.

Length of Time out of Stock

A matter which is of as much concern to stock managers as the percentage of items out of stock is the length of time for which these articles will be out of stock. Yet this latter point is rarely mentioned in the literature. If this time out of stock is calculated theoretically another interesting apparent statistical paradox arises.

Let $q/s = y$ and w be the average time out of stock.

Then $$p_0 w = \int_{-\infty}^{0} y f(y)\mathrm{d}y$$

$$= \frac{m}{\sigma_y}\int_{-\infty}^{0} y \exp\{-(y-t)^2/2\sigma_y^2\}\mathrm{d}y \quad \text{where } m = 1/\sqrt{(2\pi)}$$

For convenience let $\sigma_y = \sigma$ and $(y-t)/\sigma = v$

Then $y = t + \sigma v$ and $\mathrm{d}y = \sigma \mathrm{d}v$

When $y = -\infty \quad v = -\infty; \quad y = 0 \quad v = -t/\sigma$

Then $$p_0 w = \frac{m}{\sigma}\int_{-\infty}^{-t/\sigma}(t+\sigma v)\exp\{-\tfrac{1}{2}v^2\}\sigma \mathrm{d}v$$

$$= mt\int_{-\infty}^{-t/\sigma}\exp\{-\tfrac{1}{2}v^2\}\mathrm{d}v + m\sigma\int_{-\infty}^{-t/\sigma} v\ \exp\{-\tfrac{1}{2}v^2\}\mathrm{d}v$$

$$= t\mathrm{F}\left(-\frac{t}{\sigma}\right) - \left[m\sigma\ \exp\{-\tfrac{1}{2}v^2\}\right]_{-\infty}^{-t/\sigma}$$

where F is the error function

$$= t\mathrm{F}\left(-\frac{t}{\sigma}\right) - m\sigma\exp\left\{-\tfrac{1}{2}\left(\frac{t}{\sigma}\right)^2\right\}$$

Since $$p_0 = \int_{-\infty}^{0}\mathrm{f}(y)\mathrm{d}y = \mathrm{F}_y(0) = \mathrm{F}\left(-\frac{t}{\sigma}\right)$$

$$w = t - \frac{\sigma}{\sqrt{(2\pi)}}\cdot\frac{1}{\exp\{\frac{1}{2}(t/\sigma)^2\}}\cdot\frac{1}{\mathrm{F}(-t/\sigma)}$$

Computing these values for the example already given and with $\bar{l}=3$ and 4 and $t=1, 2, 3$, we get Table VI.

TABLE VI

Average time out of stock in months

$\bar{l}$	t 1	2	3
3	0·96	0·97	1·04
4	1·20	1·22	1·25

This gave the apparently paradoxical result that the average time out of stock *increased* with the amount of buffer stock carried. The reason for this is, of course, that the normal distribution is assumed to extend to infinity at each of its tails and thus the smaller the area under the tail being considered the further away from the mean of the distribution is the mean of this area. It can be argued that the average time out of stock is not the appropriate measure. It may be more useful to consider a 'nuisance parameter' as being the product of the number of items out of stock and the time out of stock. This is evaluated by multiplying the relevant figures from Tables IV and VI and the following values in item-months per 100 items are obtained.

TABLE VII

'Nuisance value' in item-months per 100 items

$\bar{l}$	t 1	2	3
3	3·35	1·94	1·35
4	5·52	3·66	2·50

This resolves the apparent paradox because the 'nuisance parameter' reduces with the amount of buffer stock.

In practice the paradox probably does not arise because the distribution of q/s does not extend to infinity in reality but is truncated. This can be done, for instance, by truncating the distribution when the area under the tail is reduced to $2\frac{1}{2}\%$.

Hence by integrating from $-t/\sigma$ to $-1{\cdot}96$ instead of to $-\infty$

$$p_0 w' = t\left[\mathrm{F}\left(-\frac{t}{\sigma}\right) - 0{\cdot}025\right] - \frac{\sigma}{\sqrt{(2\pi)}}\left[\exp\left\{-\tfrac{1}{2}\left(\frac{t}{\sigma}\right)^2\right\} - \exp\left\{-\tfrac{1}{2}(1{\cdot}96)^2\right\}\right]$$

$$w' = t - \frac{\sigma}{\mathrm{F}'(y)\sqrt{(2\pi)}}\left[\exp\left\{-\tfrac{1}{2}\left(\frac{t}{\sigma}\right)^2\right\} - \exp\left\{-\tfrac{1}{2}(1{\cdot}96)^2\right\}\right]$$

where $\mathrm{F}'(y)$ is the proportion of items out of stock excluding the outlying $2\frac{1}{2}\%$ area.

The results for $\bar{l}=3$, $t=1$, 2, 3 then become

TABLE VIII

Time out of stock with truncated distribution

t	Average Time out of Stock in Months
1	0·79
2	0·67
3	0·58

The Economic Batch Quantity

Once the buffer stock level has been fixed, using the appropriate criteria as discussed above, the size of an individual order can now be determined. The classical approach, originated by Camp,[3] is to order the quantity with the minimum total cost per piece by adding the costs which decline as batch size increases (set-up charges, etc.) to the costs which increase as batch size increases (storage and investment charges) and determining the order quantity at which the total is a minimum. This is illustrated in Fig. 5.

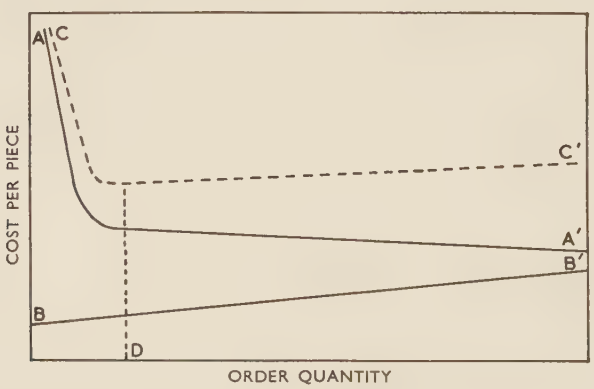


Fig. 5. Theoretical model for determination of optimum order quantity.

Curve AA′ is the curve showing how unit cost decreases with order quantity. The straight line BB′ shows how unit storage and investment

cost increases with order quantity because, at a given rate of sale, the items remain in store for a longer period and thus take up space and tie up capital for a longer period. The dotted curve CC′ is the sum of AA′ and BB′ and represents the change in the total cost per item with order quantity. It is seen that there is a minimum total cost per item at order quantity D and this is postulated as the quantity which should be ordered because it gives minimum cost and hence maximum profit.

D can be calculated as follows:

Let the equation of line AA′ be

$$\text{Cost per piece} = \frac{c}{x} + m$$

where c is the fixed charge per order, be it set-up charge or cost of placing an order or similar cost, x is the order quantity, and m is the variable manufacturing cost per item.

Let the equation of line BB′ be

$$\text{Cost per piece} = \left(\frac{c}{x} + m\right) \times \frac{x}{2s} \times \frac{P}{100}$$

where s is the rate of sales in appropriate units and P is the investment and storage costs per cent; $x/2s$ is the average length of time an individual piece remains in store and its investment and storage costs per unit time are

$$\left(\frac{c}{x} + m\right)\frac{P}{100}$$

The total cost per item, y, is therefore represented by the equation

$$y = \frac{c}{x} + m + \left(\frac{c}{x} + m\right)\frac{x}{2s} \cdot \frac{P}{100}$$

Differentiating with respect to x, equating to zero, and testing the second differential to ensure that the point under consideration is a minimum and not a maximum, we get

$$\frac{dy}{dx} = -\frac{c}{x^2} + \frac{mP}{200s}$$

$$\therefore x = \sqrt{\frac{200cs}{mP}}$$

This is the well-known square-root formula and it, and its derivatives, are bandied about to a dangerous extent. It is safe to say that any ordering scheme which uses this formula will cause serious production difficulties in circumstances which will be discussed later, and will almost certainly result in the storage of an inflated level of stock.

The economic justification for using the square-root formula is that by minimising costs it maximises profits. This justification needs closer examination. At first sight it seems appropriate because maximum profit is made from investments when the marginal return on capital is equal to the marginal cost of capital and, as shown below, this equation gives the same square-root formula as before.

$$\text{Cost per unit} = \frac{c}{x} + m$$

$$\text{Cost per batch} = c + mx$$

$$\text{Average capital investment} = (c + mx)/2$$

Let V = selling price

Then

$$\text{Gross profit per unit} = V - \left(\frac{c}{x} + m\right)$$

$$\text{Gross profit per annum} = s\left[V - \left(\frac{c}{x} + m\right)\right]$$

$$\frac{\partial \text{ Profit}}{\partial x} = \frac{sc}{x^2}$$

$$\frac{\partial \text{ Capital}}{\partial x} = \frac{m}{2}$$

$$\therefore \frac{\partial \text{ Profit}}{\partial \text{ Capital}} = \text{marginal rate of return on capital}$$

$$= \frac{sc}{x^2} \Big/ \frac{m}{2} = \frac{2sc}{mx^2}$$

For maximum profits,

$$\text{marginal return on capital} = \text{marginal cost of capital}$$

$$= P/100$$

$$\therefore \frac{2sc}{mx^2} = \frac{P}{100}$$

$$x = \sqrt{\frac{200cs}{mP}} \quad \text{as before}$$

The maximisation of profit is reasonable for a firm with unlimited capital, but for a Company with limited capital it should be seriously considered whether it might be better to work in the region where the marginal return is appreciably higher than the marginal cost and to place further investments in another product with a similarly high marginal return.

The very small advantage accruing to the Company by increasing its batch sizes beyond, say, half the quantity predicted by the square-root formula can be shown in the following example:

Let $s = 100$ per annum
$m = £2$
$c = £10$
$P = 10\%$
$V = £4$

Then $$x = \sqrt{\left(\frac{200 \times 10 \times 100}{2 \times 10}\right)} = 100$$

The net annual profit is
average investment × (average return on capital − average cost of capital)
The average investment (£) $= \frac{1}{2}(10 + 2x) = (5 + x)$

The average rate of return on capital (£)

$$= \frac{s\left[V - \left(\frac{c}{x} + m\right)\right]}{\frac{c + mx}{2}}$$

$$= 200\left(\frac{4}{10 + 2x} - \frac{1}{x}\right) = \frac{400}{5 + x} - \frac{200}{x}$$

The average cost of capital = 10%

Thus the net annual profit (£) $= (5 + x)\left(\frac{400}{5 + x} - \frac{200}{x} - \frac{1}{10}\right)$

$$= 400 - \frac{200(5 + x)}{x} - \frac{5 + x}{10}$$

$$= 100\left(4 - 2 - \frac{1}{200} - \frac{10}{x} - \frac{x}{1000}\right)$$

(*a*) With batch size $x = 100$
average investment = £105
net annual profit = £179·5

(*b*) With batch size $x = 50$
average investment = £55
net annual profit = £174·5

Thus by using a batch size of 100 in place of a batch size of 50 the capital employed is about doubled with a resulting increase in annual profit of only 3%. Most Boards of Directors could achieve a better rate of return if they doubled the amount of capital employed in their business.

The reason for this meagre result is that with a batch size of 50, £55 of capital earn a marginal rate of return of $1000/x^2 = 1000/2500 = 40\%$, whereas with a batch size of 100 the £105 of capital employed only earn a marginal rate of return of $1000/10000 = 10\%$ (as is to be

expected since this equals the marginal cost of capital of 10%).

If there are other ways within the firm where a marginal rate of return greater than 10% can be earned, then the batch size should be reduced below 100 so as to free capital for this use.

Thus an uninformed use of the square-root formula can lead to batch sizes, and hence stocks, in excess of those which are in the true interest of the firm.

There are many other reasons why the square-root formula should not be used for fixing batch sizes. Many of the assumptions about set-up charges etc. underlying its construction just do not apply in practice. This has been forcefully pointed out by Burbidge.[1]

When used to determine batch sizes in a factory, the square-root formula leads to production difficulties, as pointed out by Salveson.[4] According to this formula, each item is ordered on the factory in proportion to the square root of its own sales rate. However, each product will continue to be sold in direct proportion to its own sales rate. There will thus be a lack of balance between the order quantities of these different products and their rates of sale. This will not matter if the several articles represent only a small proportion of the plant's output because then they will not get in one another's way. But if they do account for the bulk of the plant capacity then there will be occasions when two separate articles will need to be produced at the same time. This is impossible, of course, and the makeshifts that will have to be introduced will cause departures from what was considered to be an optimum policy.

The square-root formula should no longer be advocated as a tool in solving stock control problems.

The Acceleration Effect

In the preceding sections it has been shown that the rules published in the literature for determining buffer stocks and re-order quantities are inappropriate and if applied as they are intended to be applied would lead, and indeed in many industries may have led, to a gross inflation of stocks above the necessary level.

What has not been challenged is the fundamental concept of a buffer stock level and a re-order quantity as illustrated in Figs. 1 and 2 and shown again in Fig. 6.

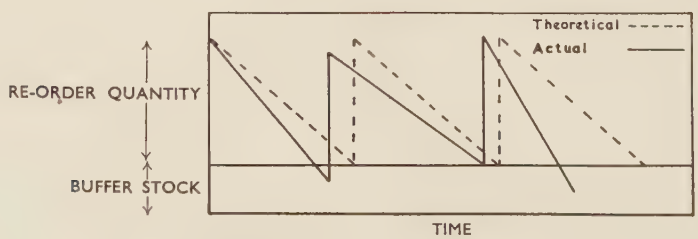


Fig. 6. Re-ordering system with buffer stock.

It is necessary to examine this system a little more closely than has been done in the literature, however, because it possesses certain properties which have not hitherto been discussed.

If sales increase steadily over a long period, the buffer stock is reached more frequently and the order quantity increases each time an order is placed. The combined effect of these two changes is to increase the total orders placed by more than the increase in sales. If the square-root formula is being used then as sales increase the order volume increases by the rate of sales increase to the power 3/2. If orders are placed in direct proportion to sales then the power is 2.

When sales decline again the opposite effect occurs. The result is to accelerate the effect of a change in sales on factory orders as illustrated in Fig. 7.

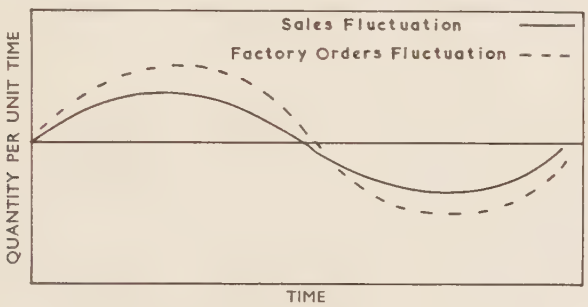


Fig. 7. Acceleration effect of sales on orders.

This acceleration effect may seem undesirable and if the goods are being produced in a factory with little scope for either increase or contraction of output without considerable dislocation then it clearly is undesirable. Paradoxically it is, nevertheless, the right thing to do in a situation where sales are showing a steady trend one way or the other. This is because orders must always be placed in advance of sales and if sales are continually increasing (say) then the orders must be increasing at a faster rate in order that the goods will be in stock when required. This out of phase effect is illustrated in Fig. 8.

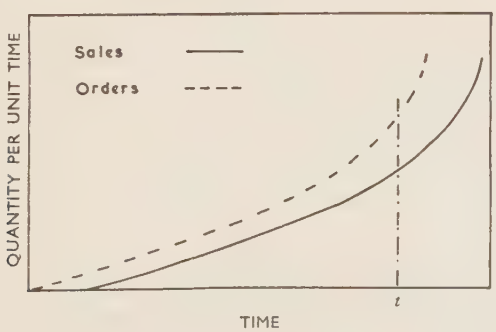


Fig. 8. Acceleration effect of sales on orders with steady trend.

The orders graph and sales graph are assumed identical in shape, as they should be if the order policy is sensible, but because the order graph must be displaced ahead of the sales graph, i.e. earlier in time, its gradient at any given point of time, say t, must be higher than that on the sales graph at the same time and thus the rate of increase of orders must be greater than the rate of increase of sales.

If a Company expects its sales to increase fairly rapidly it must therefore be prepared to produce goods at an even faster rate of increase. If it expects its sales to fluctuate over long periods then either it must be prepared for an exaggerated fluctuation in its work load or it must carry a strategic buffer stock to enable it to sell at a higher rate than it is producing, for as long as the increase of sales persists. This stock can be replenished during the period of low sales.

The important point is to be aware of the acceleration effect and to plan to meet it if possible. This has been discussed in more detail by Forrester.[5]

Conclusions

This paper has been critical of the classical statistical and mathematical methods for tackling stock control problems. The following section is aimed at putting something constructive in their stead.

In the first place it should be clearly recognised in any firm that the problem of stock levels is a problem at Managing Director level. The evidence is that far too much capital is tied up in stocks and the basic reason for this is that decisions about buffer stock, re-order quantities, etc. are too often made at lower levels in the executive chain. The result is that the firm is given too great a level of protection because individual stock controllers wish to minimise as far as possible the risk that they will be blamed if the Company should be out of stock of any commodity. This application of the 'minimax' principle in stock control has been illustrated by Mills.[6]

The Managing Director, or a properly delegated authority, should determine the maximum investment the Company can permit in stocks. The industrial statisticians can help considerably here by pointing out what different investments mean in terms of risks of stock-out, level of orders on the factory, the effects of changes in the rates of sale, and so on. As a result of the information presented to the Managing Director, he can then make a choice.

Once this choice is made, the industrial statistician then has a further part to play in setting up policies that are internally consistent. Almost everyone connected with stock control problems will know that the principal nightmare is inconsistency. Stocks of one part can be sky-rocketing while panic measures are resorted to for obtaining enough supplies of another part.

The industrial statisticians must devise rules for buffer stocks and re-order quantities for each part which provide the required level of protection, ensure that the ordered quantities are consistent with the

production requirements and total capital available, and are simple to operate and administratively convenient to check. It is a big task and one which needs close study of every individual application. The cybernetic black box principle may often be found of greater value than the results of detailed statistical analysis. An illustration of this principle applied to stock control is given by Beer.[7]

A further danger to avoid is that of the tail wagging the dog. In most factories only a few parts make up the bulk of sales. In one case known to me 25% of the items were found to constitute 75% of the sales, as shown in Fig. 9. By concentrating on rules suitable for these fast-selling parts the major part of problem was tackled and the idiosyncrasies of the remaining items could be largely ignored.

The final rule of the statistician is to arrange for adequate feed-back data to be available on the system so that its performance can be kept continually under review and appropriate changes made whenever necessary.

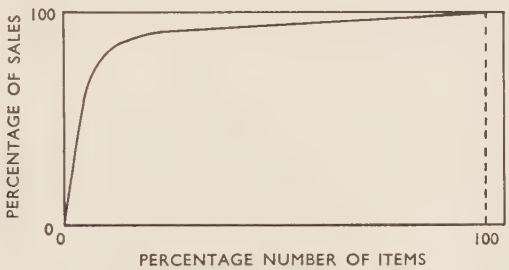


FIG. 9. Example of relationship between percentage of sales and percentage number of items when sales are highly concentrated on a minority of items.

Acknowledgements

Thanks are due to S. Babik for much of the computation and to him and the referees for their very useful critical comments on the paper.

REFERENCES

[1] BURBIDGE, J. L. (1958). 'A new approach to production control'. J. Instn. Prod. Engrs., **37**, 288.

[2] GANI, J. (1957). 'Problems in the probability theory of storage systems', *J. R. Statist. Soc.*, B, **19**, 181.

[3] RAYMOND, F. E. (1931). *Quantity and Economy in Manufactures.* McGraw-Hill, New York.

[4] SALVESON, M. E. (1956). 'A problem in optimal machine loading', *Management Science*, **2**, 247.

[5] FORRESTER, J. W. (1958). 'Industrial dynamics', *Harvard Business Review*, July-August, 37.

[6] MILLS, E. S. (1957). 'The theory of inventory decisions', *Econometrica*, **25**, 222.

[7] BEER, S. (1959). *Cybernetics and Management.* English Universities Press, London.

CLASSIFICATION PROCEDURES BASED ON BAYES'S FORMULA

A. BIRNBAUM* and A. E. MAXWELL
Institute of Mathematical Sciences, New York University
and Institute of Psychiatry, University of London

In this article Dr Birnbaum and Dr Maxwell describe some methods of classification based on the classical probability formula of Bayes, and on some of its applications suggested by the theory of statistical decision functions. The classification problems considered are general, utilising no restrictive assumptions about possible distribution forms, but assuming that distributions can be estimated from large samples of data. The method is illustrated by an exploratory application to preliminary diagnostic classification, using data on 772 patients of a psychiatric hospital.

Introduction: Classification Problems

We are concerned here with statistical problems of a kind usually designated by the term 'classification', but sometimes designated by broad usages of the terms 'estimation', 'prediction', or 'correlation'. These problems have the following form: we consider some number k of categories from each of which random observations may be obtained; thus each category is a statistical population. For example, each category may consist of individual persons, from whom a sample of persons may be selected randomly; for these individuals, the observations obtained consist of some specified characteristics or measurements or 'scores'—we shall use the term 'measurement' in its broadest sense to designate whatever specified qualitative or quantitative information constitutes the complete 'observation' obtained for one individual.

The central problem is to infer, on the basis of the measurement on an individual, the category to which the individual belongs. Thus it is the category which is to be 'predicted' or 'estimated' from the observation; in some contexts the category is referred to as the 'criterion' which is to be predicted. The categories may be strata or subpopulations of some specified population. Any procedure or rule for inferring, on the basis of the measurement on an individual, the category to which he belongs, may be represented formally by an *inference function* (which in some contexts will be called a predictor, or estimator, or a decision function). If we introduce u as a generic symbol for any one of the values which may be obtained by making a measurement of a given kind, any inference function may be represented by $\mathrm{d}(u)$: for each possible measurement u, the given inference function designates a certain category, namely category number

* Research supported by School of Aviation Medicine, USAF.

$i = \mathrm{d}(u)$; thus the values assumed by $\mathrm{d}(u)$, for various u's, are the numbers (or names, or other designating labels) of the various categories. It is sometimes convenient to represent an inference function simply by listing the u's in some convenient order, with each u followed by its corresponding i.

An important aspect of this problem is the *evaluation of* various alternative *inference functions* $\mathrm{d}(u)$ defined for a given kind of measurement u, to provide a basis for selecting the inference function to be used in a given situation. A more inclusive problem is the *evaluation of* the usefulness of various possible *kinds of measurement* for a given classification problem, to provide a basis for selecting one of them for use.

In the problem discussed in detail on pp. 163ff., the combined categories constitute the patients admitted to a mental hospital during a certain period; the individual categories are defined by alternative diagnoses of the patients, namely, (1) neurosis, (2) schizophrenia, (3) affective disorder, or (4) abnormality of personality. The diagnoses of admitted patients are not generally made until some time after admission, but upon admission certain more specific observations are available; in the example quoted these are represented by nine different traits, whose presence or absence is noted for each patient at the time of admission. The resulting 'measurement' places each patient in one of the $2^9 = 512$ logically possible patterns of traits. It is of interest to consider the problem of predicting the future diagnosis of the patient from such a measurement: that is, the problem of devising a rule for making preliminary diagnoses on the basis of such measurements alone. The reader who prefers to begin with a detailed discussion of an example, rather than general discussion of methods, may proceed from this point directly to p. 163, possibly referring back to earlier sections for supplementary discussion of the method applied there.

Methods of classification may be discussed generally in the following terms: let k denote the number of categories, and let i and j be generic labels for the categories, $i, j = 1, 2, \ldots, k$ (i is the main label; the reason for the alternative label j will be clear from the next paragraph); let m denote the number of distinct values which are logically possible for the kind of measurement under consideration, and let u be a generic label for one of these possible values, $u = 1, 2, \ldots, m$. (The restrictive assumption that the measurements are of a kind which can take only a finite number of different values can be relaxed, but to avoid technical complication the present discussion is restricted to the finite case. The infinite case becomes appreciably different only in connection with the problems of sampling discussed on pp. 162-3.) Let $\mathrm{p}(u|i)$ denote the probability that an individual will have measurement u if he is randomly selected from category number i; if each population consists of a finite number of individuals, then $\mathrm{p}(u|i)$ is also the proportion of individuals in category i who have measurement u, for each value of i and u. Each possible inference function $\mathrm{d}(u)$ takes one of the k values i for each of the m values of its argument u.

When a given kind of measurement has been adopted, the merits of any particular inference function $d(u)$ are described exhaustively (for the purposes considered in this paper) by the probabilities of errors of the various possible sorts which its use entails. When a specified $d(u)$ is used, let $r(j|i)$, for each pair i, j, denote the probability that, for an individual randomly selected from category i, $d(u)$ will take the value j (i.e. will lead to the inference that the individual is from category j). An ideal solution of a classification problem would be an inference function $d(u)$ having all error-probabilities $r(j|i)=0$, $j\neq i$; but in the problems which are of serious interest, such ideal solutions do not exist. Questions of interest are of two types:

(1) With a specified kind of measurement, and any of the possible inference functions $d(u)$, what are the attainable sets of error-probabilities $r(j|i)$, $i\neq j$? What considerations and techniques may lead to the selection of an appropriate one of the inference functions for use in a given application?

(2) Among the various kinds of measurement which might be used, what considerations and techniques may lead to the selection of an

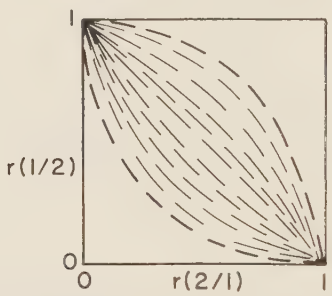


Fig. 1. Error-probability patterns with two categories and various possible inference functions (schematic diagram).

appropriate one for use in a given application? There are many considerations relevant to such choices other than the error-probabilities attainable by the use of each kind of measurement. These include convenience, costs, purposes, and perhaps the relative frequencies with which individuals from the various categories may be expected to be encountered in a given situation of application. These other considerations will not be discussed here (these aspects of classification problems have been stressed in an interesting recent study[2]); but it is important to note that all such considerations must be taken into account, jointly with the error-probabilities attainable, in any practical appraisal of alternative kinds of measurement in classification problems.

The preceding notions may be further developed and illustrated as follows; for simplicity, the discussion at this point (but not in what follows) will be formulated in terms of the special case of two categories. If $k=2$, just two sorts of errors are possible: $r(1|2)$ is the probability that an individual randomly selected from category 2 will,

when a specified $d(u)$ is used, be inferred to be from category 1; and $r(2|1)$ is the probability that an individual from category 1 will be inferred to be from category 2. If we consider one kind of measurement, and all possible inference functions $d(u)$, we obtain for each $d(u)$ a pair of error-probabilities $r(1|2)$ and $r(2|1)$. As in the schematic Fig. 1, the latter probabilities may be used as the coordinates of a point which represents a given $d(u)$; the set of points representing the various possible inference functions typically fills out some convex region within the unit square. The points along the lower-left boundary of the shaded region represent inference functions none of which can be strictly improved upon: in comparison with any such point, each possible alternative point in the shaded region represents an inference function which is worse in at least one of its error-probabilities. On the other hand, as between any two points on this lower-left boundary, there is no simple and categorical basis for choosing the 'better' one, since each of the two points is better with respect to one of its error-probabilities. Each of the latter boundary points, and their corres-

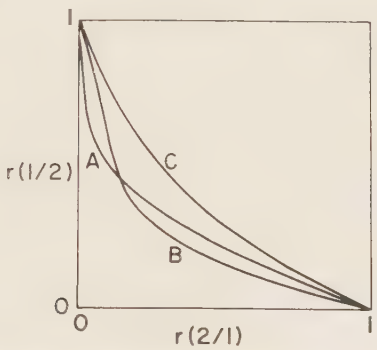


FIG. 2. Error-probability patterns for admissible inference functions based on three different kinds of measurement A, B, and C, with two categories (schematic diagram).

ponding inference functions, may be called *admissible* for reasonable further consideration for possible use; whereas other points, and their corresponding inference functions, may be called *inadmissible* for further reasonable consideration, since each of these can be strictly improved upon (i.e. one error-probability can be decreased without an increase in the other) by choosing instead a point below and/or to the left on the boundary of the shaded region.

The questions (1) above reduce formally to the problem of determining those sets of attainable error-probabilities (and corresponding inference functions) which are admissible, and to the problem of making a reasonable choice among these.

The questions (2) above may be approached as follows: suppose that three alternative kinds of measurement, A, B, and C, are available for use in a given classification problem. Each of these may be represented by a graph representing the admissible inference functions based on it, as in the schematic Fig. 2; each such graph corresponds to the lower-

left boundary of the shaded region in a figure analogous to Fig. 1. Such graphs would be expected to arise, for example, if C denotes a measurement consisting of the response-pattern on a test consisting of six dichotomous items (e.g. a measurement like that used in the example on p. 163, but based on six rather than nine traits); if A denotes a measurement like that of this example consisting of the response-pattern based on nine such items, including the six items underlying C; and if B denotes a similar measurement based on nine items, those underlying C and another three items not included in A.

In this comparison, the measurement C appears inadmissible for further consideration provided that A or B can be adopted without incurring relative disadvantages (such as costs or inconvenience) which would outweigh the advantages over C with respect to error-probabilities; if A and B are much more expensive to use than C, then of course C might reasonably be preferred. No such simple comparison is possible between A and B; in fact, if other considerations such as costs leave A and B equally attractive, then it is clear that one might prefer either A or B, depending on which particular point on the A or B curves in Fig. 2 one chose to attain.

If in a given classification problem we find that a specified kind of measurement gives no inference function whose various error-probabilities are suitably small, we are led to conclude that the *kind* of measurement being used is not sufficiently informative, and we are led to seek an alternative kind of measurement which may prove more satisfactory. On the other hand, if one kind of measurement leads to inference functions whose error-probabilities are considerably smaller than is necessary in some application, we may wish to consider another kind of measurement, less expensive or easier to use, which would not increase the error-probabilities intolerably.

Finally, on the basis of one or more kinds of measurement, there may be several inference functions available, each of which has error-probabilities sufficiently small to be tolerated in a given application.

Considerations like those described above have helped us up to this point in simplifying the problems of choice among inference functions and kinds of measurement, but they leave unresolved the problem of making a final choice from the many admissible inference functions remaining under consideration. Since the various sorts of possible errors are generally of qualitatively different kinds, there appears to be no requirement of reasonableness which compellingly leads us to seek some formal rule by which to make such selections. It could perhaps be postulated that each of the possible sorts of errors has a 'cost', and that these costs could be expressed on some single numerical scale of utility. This may be the case in some applications, but in many others, such as the later example, the possibility of measuring these costs in terms of some single unit of utility seems quite hypothetical (except for limited administrative purposes). Any attempt to apply such a formulation even as a formal step toward choosing an

inference function would seem to risk distorting or obscuring some of the practical judgements and value-judgements appropriate to the specified situation of application. Such judgements can well be brought to bear informally in direct consideration of the merits of various rules as represented simply by their error-probabilities.

There is an extensive literature and body of theory devoted to problems of inference or decision-making formulated in terms of costs and utilities associated with the various possible errors (see, for example, reference 5 and references therein). In the statistical theory of decision functions initiated by Wald,[8] inference problems are formulated by the introduction of loss-functions which represent, in terms of some common unit of utility, the costs of the various possible sorts of errors. It seems unfortunate that much writing in the field of statistical decision theory fails to mention the fact that, even when such loss-functions are quite hypothetical from the standpoint of the situation of application, nevertheless the inference-functions (or decision-functions) determined by their formal use are typically admissible when they are appraised more directly and simply in terms of the probabilities of errors of various sorts to which their use leads. A notable exception is the paper by Lindley[4]; parts of the present paper may be regarded as informal exposition of parts of Lindley's paper.

Construction of Admissible Inference Functions

A general method for constructing various admissible inference functions is based on the introduction, in a *hypothetical formal* way, of possible *a priori* probabilities for the k categories of interest. Let us *imagine* that there exists a population which consists of our k categories as sub-populations or strata, combined in the proportions g_i, for $i=1, 2, \ldots k$; in other words, let us imagine that there is a composite population such that if we randomly select an individual from it, the probability is g_i that the selected individual belongs to category i, $i=1, 2, \ldots, k$. Among all individuals within a given category i, we have, as before, that the measurements u of some specified kind have the probability distribution function $\mathrm{p}(u|i)$. It then follows formally from the formula of Bayes in the elementary theory of probability that, if an individual is randomly selected from this hypothetical composite population and if only his measurement u is observed, then the probability that he belongs to category number i is given by

$$\mathrm{p}(i|u) = g_i \mathrm{p}(u|i)/\mathrm{p}(u)$$

Here $\mathrm{p}(u)$ is the probability that any individual randomly selected from the hypothetical composite population has measurement u, i.e. $\mathrm{p}(u) = \sum_{i=1}^{k} g_i \mathrm{p}(u|i)$. An important simplification, which is useful in many applications, is obtained by noting that the *relative* magnitudes of the $\mathrm{p}(i|u)$'s, for a given measurement u, are available without computation

of $p(u)$, since $p(i|u)$ is simply proportional to $g_i p(u|i)$. In this hypothetical context, Bayes's formula suggests the inference that, if an individual is observed to have measurement u, then he is most likely to belong to that category i for which the quantity $g_i p(u|i)$ is largest. This Bayesian rule of inference may be represented by an inference function $d(u)$ defined as follows:

For each u, let
$d(u) =$ that number i which maximises $g_i p(u|i)$.

(If two or more i's give the same maximum value, the arbitrary choice of one of them, or the definition of a random method for selecting one of them each time such a u is observed, leads to a definition of a particular Bayesian rule.) It is important to note that the definition of each Bayesian inference function $d(u)$ makes use of constants g_i which may be *completely arbitrary* (*and hypothetical, as probabilities*). Once such a $d(u)$ is defined, however, we can immediately compute and consider the various error-probabilities $r(j|i)$ which will result from the use of $d(u)$. *Each such Bayesian* $d(u)$ *is recommended for serious consideration by its fundamental property of admissibility, namely that there is no alternative* $d(u)$ *which gives smaller values to any of the error-probabilities* $r(j|i)$ *except at the price of increasing some other error-probabilities. This property is not at all affected by the arbitrary or hypothetical nature of the constants* g_i *used.*

This fundamental property of Bayesian inference rules has taken a central place in the development of the theory and techniques of mathematical statistics, beginning with its use by Wald in his theory of statistical decision functions. However, it seems not yet widely appreciated that the method of demonstration of this property is essentially an elementary one: in the hypothetical composite population having probabilities g_i for the various categories, the problem of classifying a randomly selected individual on the basis of a measurement u, so as to maximise the probability of correct classifications, is a simple problem of classical probability theory. Its solution is given by the Bayesian rule represented by the $d(u)$ defined above, and the resulting minimised probability of misclassifications is given by

$$\sum_{i=1}^{m} g_i \sum_{j \neq i} r(j|i)$$

where the second summation symbol $\sum_{j \neq i}$ denotes taking a sum of all the error-probabilities $r(j|i)$ for a given i, $j \neq i$. If such a $d(u)$ were not admissible, there would exist another inference rule having none of its error-probabilities larger, and one or more of them smaller, than the $r(j|i)$'s of $d(u)$ in the above expression. But this would give a smaller value to the above expression for probability of misclassifications, contradicting the fact that $d(u)$ minimises this probability. This contradiction proves that $d(u)$ is admissible—a property not compromised by the fact that the g_i's are arbitrary or hypothetical. [This

proof tacitly assumed that each g_i was positive, not zero, as will usually be the case in applications of interest. Even if some g_i's are taken equal to zero, an equally simple proof is possible provided that tied maximised values of the expression $g_i p(u|i)$ (for various i's), which is used in the definition of $d(u)$, occur with negligible probabilities: if there are no such ties, then the Bayesian $d(u)$ is determined uniquely for each u in the above definition. Hence every different inference function gives a larger probability of misclassifications, and hence every inference function different from $d(u)$ has at least one of its error-probabilities larger than the corresponding $r(j|i)$ of $d(u)$.]

The working methods which are suggested by these considerations are the following. In any classification problem, as soon as the respective distributions $p(u|i)$, of some given kind of measurement u within each category i, are known, we can immediately write the definitions of a number of admissible Bayesian inference functions $d(u)$ defined as above in terms of the relative magnitudes of the quantities $g_i p(u|i)$; each arbitrary choice of the hypothetical probabilities g_i defines one such function. In a given classification problem it will often be of interest to consider for possible use at least several different such functions $d(u)$ and to appraise and compare them (and the kind of measurement u on which they are based) on the basis of their respective sets of error-probabilities $r(j|i)$. In order to calculate the error-probabilities of any given inference function $d(u)$, we have merely to apply the formula

$$r(j|i) = \sum_{u \text{ in } A_j} p(u|i)$$

where A_j denotes the set of values of u for which $d(u) = j$. Calculations which are generally easier to carry out, and which may suffice for some purposes (especially for preliminary consideration of a given decision function) are those of the respective probabilities of correct classifications:

$$r(i|i) = \sum_{u \text{ in } A_i} p(u|i)$$

As soon as the error-probabilities of several Bayesian inference functions have been obtained, the problem of evaluating, comparing, and possibly adopting one of them can be considered further in the ways discussed and illustrated in the introductory section. If the number of categories k is larger than two, then diagrams like Figs. 1 and 2 cannot be used to represent the sets of $r(j|i)$'s of the various inference functions. But inspection of the numerical values of these $r(j|i)$'s and/or of the $r(i|i)$'s can be the basis for judgements of the sort illustrated by use of these figures. [It may sometimes be useful to sketch such figures for a given pair of categories, temporarily omitting from consideration all but two categories; in this way, for example, if k exceeds two, a figure like Fig. 1 above would result from plotting points with coordinates $r(2|1)$ and $r(1|2)$ computed for several Bayesian

$d(u)$'s; other figures might be sketched to represent the $r(i|j)$ and $r(j|i)$ values attainable for other pairs i and j of categories.]

An important practical question which must be met in any application of these methods is the choice of one or several sets of hypothetical probabilities g_i to use in defining the first Bayesian inference rules to be examined; and then at each stage of consideration of inference rules, the choice of successive new sets of g_i's to be tried when additional exploration of the class of Bayesian inference rules seems worth while. Unfortunately, there are no general rules available which give very useful quantitative information to guide these choices, and an element of trial-and-error seems inevitable in the selection of trial values of g_i's, except for the initial choice in the case where *a priori* estimates are available. However, much qualitative information becomes available as to the directions in which the g_i's should be varied to obtain desired modifications of successive inference rules considered. The basic consideration here is the property of admissibility of each of the inference rules obtained [illustrated by Figs. 1 and 2, and by similar figures or analogous tables of $r(j|i)$ values, which may be filled in progressively in a given application as successive inference rules are considered]. The knowledge that the set of points [whose coordinates are $r(j|i)$'s, $j \neq i$] which one is exploring is a convex curve or surface, is of great help. Also useful is the fact that, if a given set of g_i's gives error-probabilities which include too small a probability of correct classification for individuals from a certain category i, then by increasing only the corresponding g_i, and decreasing some or all of the other g_i's, one will generate a new inference function with smaller probabilities of errors of each kind possible for individuals from the given category. To increase probabilities of correct classifications for individuals from a specified pair of categories, one should try some increase in only the corresponding two g_i's, adjusting some or all of the others downward, and so forth. The information obtained by inspection of the error-probabilities $r(j|i)$ [and/or the probabilities $r(i|i)$ of correct classifications] for each inference rule examined will provide, along with such general qualitative considerations, more useful guidance in an application of these methods than it seems possible to provide in any more formal suggestions.

If $k > 2$, the 'simple Bayesian rules' discussed up to this point, which can be generated by varying the g_i's, do not include all admissible rules, but for many applications it will be simpler and sufficient to restrict attention to them. The additional admissible rules, which may be called 'generalised Bayesian inference rules', are distinguished from the 'simple' rules so far considered by two features:

(1) On consideration of their probabilities $r(i|i)$ of correct classifications alone, they can each be strictly improved upon by some simple Bayesian rule.

(2) They are not on that account inadmissible, since each of them has a pattern of probabilities $r(j|i)$, $j \neq i$, of incorrect classifications

which cannot be matched or improved upon by any other (simple or generalised) Bayesian rule. In exploring the class of admissible rules for a given application, it will be simplest to restrict initial consideration to simple Bayesian rules; and to consider some generalised Bayesian rules only if the particular *patterns* of $r(j|i)$'s, $j \neq i$, found among simple Bayesian rules are unsatisfactory. It may then be worth trying to modify them, even at the cost of some worsening of a set of $r(i|i)$'s attainable by a simple Bayesian rule.

The basic method of constructing generalised Bayesian rules is the following generalisation of the above method. Let (h_{ji}) be a k by k array of arbitrary positive numbers, except that $h_{ii} = 0$ for each i. For every inference rule $d(u)$, the weighted sum of its error-probabilities

$$R = \sum_{i=1}^{k} \sum_{j=1}^{k} h_{ji} r(j|i)$$

can, in principle, be computed. A generalised Bayesian inference rule is one which minimises R, for any given array (h_{ji}). It follows, by an argument similar to that above, that such a rule is admissible. Since $r(j|i) = \sum_{u \text{ in } A_j} p(u|i)$, it is easily seen that R is minimised by the inference function which takes, for each u, the value j for which

$$R_j(u) = \sum_{i=1}^{k} h_{ji} p(u|i)$$

is minimised. Thus to construct a generalised Bayesian rule, given the distributions $p(u|i)$ and an array (h_{ji}), requires the computation and comparison, for each u, of the k quantities $R_j(u)$, as contrasted with the k quantities $g_i p(u|i)$ required for a simple Bayesian rule.

In exploring the class of such rules for a given application, it is useful to note that increasing one h_{ji}, $j \neq i$, while leaving the others unchanged, tends to reduce the corresponding $r(j|i)$ and of course to increase one or more other error-probabilities.

(The h_{ji}'s can be taken as non-negative, rather than *positive*, subject to qualifications; cf. reference 4 for a precise theoretical discussion.)

Interpretation of Admissible Inference Functions in Research

The preceding discussion has centred on the problems of finding kinds of measurement, and inference functions, which minimise error-probabilities $r(j|i)$. In some contexts of application, the principal interest in the quantities $r(j|i)$ will be in their direct practical interpretations as probabilities of error. In other contexts, however, they may also serve a research purpose. For example, relations between a given set of symptom-patterns (u) and a given set of diagnostic categories or diseases (i) are represented, in the form of an admissible inference function and its $r(j|i)$'s, in ways which may prove interesting and possibly suggestive for purposes of theoretical investigation; of course the choice of categories, methods of measurement, and interpretation of such results all lie in the area of responsibility of the empirical research scientist.

Relationships to Other Methods; Sampling Errors

Although the method presented here is a simple and logical one for problems of classification, it seems to have received relatively little consideration in comparison with other methods. One reason for this, no doubt, is that this method requires relatively extensive data for its application since the distribution functions $p(u|i)$, for each category i, must be known. If the method of measurement is such that the number of values which may possibly be assumed by u for a given i is large, then the precise estimation of $p(u|i)$ will require a very large sample from each category i. For this reason the method might be called an actuarial method; it will be applicable generally only where large masses of data are easily available or are accumulated from large samples of individuals in a routine way. In cases where the number of possible values of u is small, the required sample sizes will be relatively small, but they must still be large enough to give precise estimates of all $p(u|i)$'s.

The principal advantage of this method, as compared with others, is that no restrictive theoretical assumptions are imposed on the forms which the distributions $p(u|i)$ may take. To the extent that special known forms may be assumed to hold for the $p(u|i)$'s (on empirical or theoretical grounds), derivations like those above lead to admissible (Bayesian) inference functions having special forms related to those of the $p(u|i)$'s. If the $p(u|i)$'s can be assumed known except for the values of several parameters (for example, to have multivariate normal distributions, with common unknown covariance matrices, as in reference 1), then considerably smaller random samples from each category will suffice to give the required precise estimates of the $p(u|i)$'s; the refinements required to take precise account of the errors of estimation with moderate sample sizes in such cases have been investigated to a considerable extent but have not yet been completely solved. The corresponding refinements required to take precise account of errors of estimation of the $p(u|i)$'s in the present method from samples which are not very large (relative to the number of possible values of u) have not yet been developed, although some relevant research has been carried out (cf. reference 3, for example). It seems clear that such developments would require the solution of some quite difficult problems of theoretical statistics, and that the results of such investigation would not obviate the need for quite large sample sizes in typical applications of the method.

If the $p(u|i)$'s are estimated from samples which are not very large, then the formal application of the methods described above to such estimates of the $p(u|i)$'s will lead to inference functions and error-probabilities which are estimates of the unknown desired inference functions and their error-probabilities. These estimates will typically be subject to fairly large sampling errors, and to appreciable biases in the direction of underestimating both (*a*) the error-probabilities which

actually hold for the obtained inference function, and (*b*) the error-probabilities which are attainable when the $p(u|i)$'s are known or precisely estimated. There seems to be no general method now available which is useful in determining the magnitudes of such biases. Analogous questions have been investigated in a special classification problem in reference 6.

Illustrative Example

The methods outlined above were applied to a sample of 772 males, aged 16-59, who had been in-patients at the Bethlem Royal and Maudsley Hospitals, London. Information concerning the *presence* or *absence* of numerous psychiatric characteristics was readily available for these patients on Hollerith punched cards, and on the basis of this and other information the patients had been allotted to one or another of the customary psychiatric classification categories.

For our purpose nine traits were chosen which appeared to give a reasonable cross-section of patient traits and which themselves overlapped only to a small degree (the tetrachoric correlations between the traits were known from an earlier study[7]). The nine traits were:

1. Hysterical symptoms, lifelong or episodic
2. Anxiety
3. Schizophrenic type of thought disorder
4. Depression
5. Compulsive acts and/or obsessional thoughts
6. Lack of confidence when in society
7. Definite mood variations before present illness
8. Impulsiveness and/or aggressiveness
9. Hypochondriacal attitude towards illness.

Current psychiatric diagnostic categories are numerous, but for this study it was possible to group the 772 patients into four broad categories. An estimate of the *a priori* probability that a patient entering the hospitals in question would fall into a given one of these categories was obtained from hospital records for a larger group of patients than the sample of 772. The actual sample sizes N_i and the estimated *a priori* probabilities g_i' of each category are given below:

Category	N_i	g_i'
Neurotics	341	0·461
Schizophrenics	174	0·237
Affective disorders	148	0·157
Abnormal personality	109	0·145
	772	1·000

The methods described here might equally well have been applied with one or more of these broad categories sub-divided into more

refined categories. For example, the neurotic group of 341 patients is composed of

Anxiety reactives	104
Hysterics	35
Obsessionals	31
Neurotic depressives	123
Psychoneurotics and others	48
	341

However, had groups as small as these been employed the number of traits used—for reasons of sampling discussed above—should probably have been reduced considerably.

If we use nine dichotomously scored symptoms, the number m of distinct answer patterns possible is 2^9, or 512. By means of a Hollerith sorter, the 772 cards, one for each patient, were divided into two groups, depending on whether the response to the first symptom was affirmative or negative. These two groups of cards were subdivided further depending on the responses to the second symptom, and so on until all nine symptoms had been considered. This task proved much less onerous than had been expected, for of the 512 possible answer patterns only 178 occurred.

The number of patients from each diagnostic category giving each of the 178 observed answer patterns was then recorded. These numbers were expressed as proportions of the total number in the category to which they belonged, giving the required (estimated) values $\mathrm{p}(u|i)$. These proportions were each multiplied by the corresponding g_i'. Each answer pattern u was then allotted to the category i for which the weighted proportion, $g_i'\mathrm{p}(u|i)$, was largest. For example, the answer pattern

$$- \quad - \quad - \quad + \quad - \quad + \quad - \quad - \quad -$$

occurred 28 times, 13 times among neurotics, 8 times among affective disorders, 5 times among schizophrenics, and twice among abnormal personalities. Since the numbers in these categories are 341, 148, 174, and 109 respectively, the proportions of patients in the respective categories having this answer pattern are

$$0{\cdot}0381, \quad 0{\cdot}0541, \quad 0{\cdot}0287, \quad 0{\cdot}0183$$

When these proportions are multiplied by the *a priori* probabilities g_i' of the categories we obtain 0·017 573, 0·008 513, 0·006 807, 0·002 655, respectively; these numbers are proportional to the *a posteriori* probabilities that a patient with the given answer pattern will belong to the respective categories. Since the first of these weighted proportions is the largest, the answer pattern in question is allotted to the neurotic group.

Of the 178 different answer patterns observed, 81 were allotted in this manner to Neurotics (eight of them accounting for 34 per cent of the cases), 12 to Affective Disorders, 57 to Schizophrenics (eight of them accounting for 36 per cent of the cases), and 28 to Abnormal Personalities.

This completed the determination of a Bayesian rule for making preliminary diagnoses on the bases of the specified nine traits, and illustrates the general procedure for determining Bayesian inference rules $d(u)$. Such a rule can be conveniently represented simply by a listing of the answer patterns u in some convenient order, each followed by the number i of the category to which it is allotted.

It must be stressed that although in this example estimated *a priori* probabilities g_i' have been used as a starting-point in developing a rule, there is no necessity for this. Any self-consistent set of g_i values may be used to develop a rule. But where, as in this case, estimates of *a priori* probabilities are available, it is logical to use them first and to examine the consequences of the rule to which they lead.

On the assumption that the proportions $p(u|i)$ used are correct (or estimated with sufficiently high precision), then regardless of the source of the g_i's used one can say that the inference rule obtained is admissible, i.e. that any other rule having some error-probabilities smaller will necessarily have others larger. To evaluate the rule obtained, its probabilities $r(i|i)$ of correct classifications and $r(j|i)$ of incorrect classifications were computed as indicated on p. 159, and are given in Table I.

TABLE I

$r(j|i)$ = *the (conditional) probability that a patient from category i will be classified as belonging to category j; the classification rule is based on the estimated probabilities* g_i'

		i			
		1 Neurotics	2 Schizophrenics	3 Affective Disorders	4 Abnormal Personality
j	1	(0·856)	0·241	0·595	0·422
	2	0·038	(0·706)	0·034	0·073
	3	0·073	0·034	(0·338)	0·055
	4	0·032	0·017	0·034	(0·450)

The diagonal entries, in parentheses, are probabilities of correct classifications: for example, the (estimated) probability that a 'neurotic' will be correctly classified is 0·856. The (estimated) probability that a 'neurotic' will be classified as a 'schizophrenic' is 0·038; the (estimated) probability that a 'schizophrenic' will be classified as a 'neurotic' is 0·241; etc.

On this basis, the evaluation and revision of such a classification rule might proceed as follows: if the probabilities of correct classifications for individuals in categories 3 and 4 (0·338 and 0·450 respectively)

are judged intolerably low, these can be raised by constructing a new classification rule based on appropriately modified g_i values, but only at the cost of lowering one or both of the probabilities of correct classification for individuals in categories 1 and 2 (0·856 and 0·706 respectively).

If any lowering of the latter probabilities would also be intolerable, then the kind of measurement (set of nine traits) being used must be judged inherently too uninformative for the purpose intended; consideration might be given to alternative and/or more extensive sets of traits which might prove more informative. On the other hand, if some lowering of the last-mentioned probabilities of correct classification might be tolerated, consideration could be given to another classification rule, developed just as that described above except for the use of larger values of g_3 and g_4. To this end, taking each $g_i = 0{\cdot}25$, a new classification rule was determined on the basis of the same data; its probabilities of correct and incorrect classifications are given in Table II.

It might be thought illogical to use g_i values which differ markedly from the prior probabilities estimated with a fair degree of precision, as assessed by sampling error. However, it must again be stressed that the g_i values are hypothetical quantities introduced for the purpose of developing objective rules; they are a means to an end and may be adjusted to give various patterns of error-probabilities. The use of a g_i different from the estimate of the prior probability does not necessarily imply that the estimate is considered to be incorrect.

TABLE II

$r(j|i)$ = *the (conditional) probability that a patient from category i will be classified as belonging to category j; the classification rule is based on the second set of* g_i *values (each* $g_i = 0{\cdot}25$)

		i			
		1 Neurotics	2 Schizophrenics	3 Affective Disorders	4 Abnormal Personality
j	1	(0·388)	0·057	0·140	0·072
	2	0·000	(0·577)	0·008	0·000
	3	0·384	0·229	(0·772)	0·128
	4	0·228	0·137	0·080	(0·800)

The probabilities of correct classifications have been raised as intended for category 3 (from 0·338 to 0·772) and category 4 (from 0·450 to 0·800), at the cost of decreasing those for category 1 (from 0·856 to 0·388) and category 2 (from 0·706 to 0·577).

Evaluation of this second classification rule can now be made on the basis of Table II; this might proceed as follows. If the probabilities of correct classification for categories 1 and 2 are now judged too low, and if no lower values for probabilities of correct classification for categories 3 and 4 could be tolerated, then the kind of measurement

being used must be judged too uninformative, and alternative kinds of measurement considered. If the pattern of probabilities in Table II is judged satisfactory for the purpose at hand, this classification may be adopted, with the assurance that it is admissible (cannot be strictly improved upon using the present kind of measurement) as well as satisfactory. Finally, still other sets of values g_i could be used to generate alternative admissible classification rules; consideration of each such additional rule on the basis of its $r(j|i)$ values contributes to a more complete picture of the possibilities inherent in the given kind of measurement (generalising the considerations of the simple case represented in Fig. 1 above).

The validity of the quantities $r(j|i)$ given for the above classification rules, in application to new incoming patients, would depend crucially on the assumption that distributions $p(u|i)$ of response-patterns within each category i are the same for new patients as for the previous patients represented by the data used in deriving the rules. If these probabilities change, then the computed values of $r(j|i)$'s would generally no longer be correct; the actual probabilities of incorrect classifications of various kinds could then become higher, or lower, than the computed values. This validity also depends on the assumption that errors in estimation of the $p(u|i)$'s are negligible. Validity of the (conditional) probabilities $r(j|i)$ does *not* depend on the actual proportions of patients in the categories (i), nor on whether these proportions remain constant.

Another quantity of possible interest for the appraisal of a classification rule is the total probability of correct classifications. However the interpretation and use of this quantity requires additional careful qualifications. If we write g_i' to denote an *assumed true proportion* of incoming patients in category i, in contrast to g_i which has above denoted merely a convenient algebraic variable, and apply a classification rule with known $r(j|i)$'s, then the proportion of incoming patients correctly classified will be (in expected value) $C = \sum_{i=1}^{4} g_i' r(i|i)$. For example, for the first rule above (Table I), if we now assume each $g_i' = 0{\cdot}25$, we obtain

$$C = \frac{1}{4}(0{\cdot}856) + \frac{1}{4}(0{\cdot}706) + \frac{1}{4}(0{\cdot}338) + \frac{1}{4}(0{\cdot}450) = 0{\cdot}587$$

But taking the g_i' values actually estimated from the data above and used as g_i values in deriving this classification rule, we obtain similarly $C = (0{\cdot}461)\ (0{\cdot}856) + \ldots = 0{\cdot}680$. Finally, if the second classification rule (Table II) is used, and we assume the true g_i' values each to be $0{\cdot}25$, we obtain $C = \frac{1}{4}(0{\cdot}388) + \ldots = 0{\cdot}634$. (It is an algebraic fact, which was used in derivations in previous sections, that for any given values g_i', C is maximised by use of the classification rule derived as above with the variables g_i taken equal to the given values g_i'.)

The validity of such a computed value C depends, in addition to the above-mentioned requirements for validity of $r(j|i)$'s of a given rule, on the correctness of the values g_i' for the population of patients to whom the rule is currently being applied. The significance of a numerical value for C, or equivalently for the total probability of incorrect classifications with a given rule:

$$1 - C = \sum_{i=1}^{4} g_i' \sum_{j \neq i} r(j|i)$$

is limited, since the last expression is a weighted average of probabilities of quite different kinds of errors; for different purposes, the different sorts of errors will be of quite different seriousness and significance. For special purposes it might be desired to assign a numerical measure of cost to each type of error: if $W(j|i)$ denotes the cost of the type of error whose probability is $r(j|i)$, then the expected cost of classification errors, per patient, is

$$W = \sum_{i=1}^{4} g_i' \sum_{j \neq i} W(j|i) r(j|i)$$

If such purposes were of primary importance, and if it were desired to adopt a classification rule which minimised W when certain values g_i' were assumed, the desired rule could be found as indicated earlier, by identifying the h_{ji}'s there with the $g_i' W(j|i)$'s here.

When values of the quantity C are considered, another way of appraising overall the kind of measurement used is available. If it is assumed that the true proportion g_i' of Neurotics is 0·461, and that the other g_i' values are smaller, without use of response patterns it is possible to achieve the probability of correct classifications $C = 0{\cdot}461$ by use of the simple rule which classifies all patients as Neurotic, and this probability will be the maximum obtainable with this simple type of rule. The usefulness of the kind of measurement used is reflected in the increase from this number to the value $C = 0{\cdot}680$, an increase in correct classifications of 21·9%.

Acknowledgement

We are grateful to Miss N. Hemsley for carrying out the Hollerith work.

REFERENCES

[1] Anderson, T. W. (1958). *Introduction to Multivariate Statistical Analysis*, Chapter 6. John Wiley, New York.

[2] Cronbach, L. J. and Gleser, G. C. (1957). *Psychological Tests and Personnel Decisions*. University of Illinois Press, Urbana.

[3] Fix, Evelyn and Hodges, J. L. *Discriminatory Analysis; Non-parametric Discrimination.*
(*a*) *Consistency Properties.* USAF School of Aviation Medicine Project No. 21-49-004, Report No. 4, Feb. 1951.
(*b*) *Small Sample Performance.* USAF School of Aviation Medicine Project No. 21-49-004, Report No. 11, Aug. 1952.

[4] Lindley, D. V. (1953). 'Statistical inference', *J. R. Statist. Soc.*, B, **15**, 30.

[5] Luce, R. D. and Raiffa, H. (1957). *Games and Decisions.* John Wiley and Sons, New York.

[6] Stoller, D. S. (1954). 'Univariate two-population distribution-free discrimination', *J. Amer. Statist. Ass.*, **49**, 770.

[7] Trouton, D. A. and Maxwell, A. E. (1956). 'The relation between neurosis and psychosis', *Journal of Mental Science*, **102**, 1.

[8] Wald, A. (1950). *Statistical Decision Functions.* John Wiley, New York.

OPERATIONAL STATISTICS IN INSTRUMENTAL ANALYSIS

ALLAN B. CALDER
Rutherford College of Technology, Newcastle-upon-Tyne

In this article Dr Calder shows how simple statistical methods were used to study some problems arising in spectrochemical analysis. As a result, revised calibration methods were introduced with consequent saving of time and money.

Introduction

It must be stated at the outset that in chemical and instrumental analysis mathematical calculation cannot obviate the need for common sense and sound analytical technique. Nevertheless experimental time can often be economised and the reliability of the deductions increased by the use of statistical methods. This paper is concerned with the application of statistical treatment to certain problems associated with a particular type of instrumental analysis and emphasises the immense saving in operational time and costs which can be effected, without loss of reproducibility, by correct interpretation of the analytical data.

Standardisation

In instrumental methods of chemical analysis the calibration usually involves plotting the 'instrumental response' (or some function of it), for the element which it is desired to determine, against the concentration of that element. For example, emission spectrographic analytical methods are based on the assumption that

$$I_E/I_S = a_E(C_E)^{n_E}/a_S(C_S)^{n_S} \quad \ldots\ (1)$$

where I_E = intensity of a particular spectral line of the element which it is desired to determine;
I_S = intensity of a particular spectral line of an element present in known amount in the sample (this element is known as an internal standard);
C_E, C_S = respective concentrations of the elements concerned in the sample;
n_E, n_S = constants dependent on background effects and light absorption;
and a_E, a_S = constants dependent on the conditions of excitation.

Equation (1) may be rewritten in the form

$$\log(I_E/I_S) = n_E \log C_E + \log[a_E/a_S(C_S)^{n_S}] \quad \ldots\ (2)$$

The function $\log(I_E/I_S)$ is calculated from the densities of the spectral lines, d_E and d_S, which in turn are derived from galvanometer deflection values (instrumental responses), θ_E and θ_S, obtained on a

suitable photometer. It would be quite in order to plot $\log I_E$ against $\log C_E$ for the calibration, but the inclusion of I_S in the calculation serves to increase the accuracy by compensating for small variations in instrumental performance between successive analyses. Theoretically it is preferable for the analysis and internal standard lines to lie close together in the spectrum on account of the variation in the density-intensity relationship with wavelength, and only lines should be chosen whose intensities behave similarly with changes in excitation. It is not always possible in practice to fulfil this latter condition. In some cases, as for example the determination of certain alloying constituents in steel, it is possible to select as the internal standard an element which is present in the analysis material in constant amount (e.g. iron, Fe); i.e. C_S is constant. In others, such as the determination of trace elements in biological concentrates, it is not possible to find an element present in fixed amount and recourse is made to the 'variable internal standard' method, in which C_S varies. In steel analysis the calibration graph is therefore obtained as a straight line of slope n_E by plotting $\log(I_E/I_S)$ against $\log C_E$, whereas in biological analysis the calibration involves three steps: (i) obtaining plots of $\log(I_E/I_S)$ against $\log C_E$ for different values of C_S; (ii) selecting one of these curves as the 'working curve' (preferably one with a C_S value nearest the general run of the samples to be analysed); and (iii) deriving a correction graph for the C_S value of the sample. This latter method of calibration is carried out in the following manner.[1]

$\text{Log}(I_E/I_S)$ is plotted against $\log C_E$ for sets of standard mixtures. The mixtures contain increments of analysis elements (10-3162 p.p.m.) in matrices similar to the samples to be analysed, but with increments of the internal standard, which is usually iron, from one set to the next. Duplicate spectra are recorded and typical series of working curves are obtained as parallel straight lines of slope $45° \pm 0{\cdot}5°$ (i.e. n_E = unity) from 10-1000 p.p.m. when the appropriate scales, namely one $\log(I_E/I_S)$ unit = one log concentration unit, are chosen. Above 1000 p.p.m. concentration the slopes of the curves fall off gradually owing to absorption in the light source, but the vertical displacements between the curves are fairly constant over the 1000-3162 p.p.m. range. The mean value for the vertical displacement of a working curve due to change in internal standard content is obtained by measuring the displacement at close intervals[2] (say at 1 cm horizontal distances if the log concentration scale from 10-3162 p.p.m. covers 25 cm) and averaging. The values obtained are plotted against the percentage C_S to provide a correction curve. It is convenient if the measurements are made from the working curve based on a C_S value nearest the general run of samples to be analysed (in this case 5·0% Fe_2O_3). The correction for 5·0% Fe_2O_3 is then zero, while that for lower contents is negative and that for higher contents is positive. From the correction curve, a value for the correction for any internal standard content (determined by a standard chemical procedure) can be read

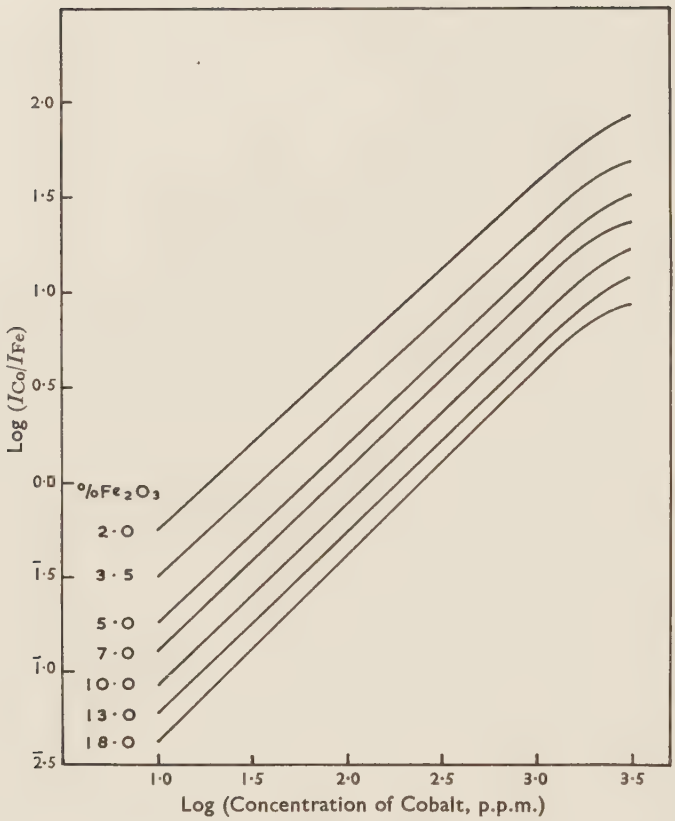


Fig. 1. Calibration curves for cobalt at various percentages of the internal standard, Fe_2O_3.

between the limits for which working curves were obtained; after this curve has been constructed, only the standard working curve (i.e. 5·0% Fe_2O_3) is retained. Fig. 1 shows a series of working curves obtained in a typical calibration for cobalt and Fig. 2 the derived correction

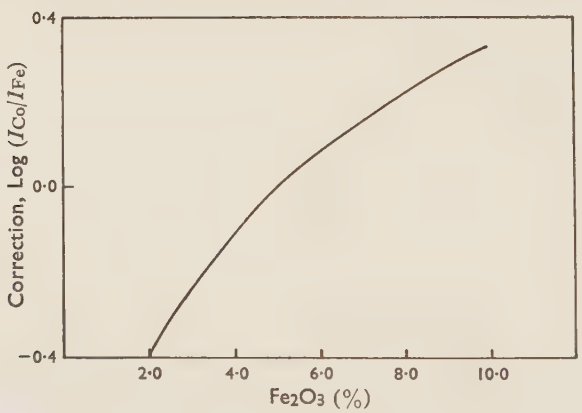


Fig. 2. Correction curve for cobalt.

graph. The procedure with subsequent measurements of $\log(I_E/I_S)$ is then to add the correction factor for the appropriate value of C_S (read from Fig. 2) and to determine $\log C_E$ from the working curve, using the corrected value of $\log (I_E/I_S)$.

This method has certain disadvantages: (*a*) the entire procedure is time-consuming; (*b*) the photometry and curve-drawing soon become tedious, especially when calibration for more than one element is contemplated; (*c*) continuous photometry over several hours imposes a strain on the photometer photo-cells (as well as on the operator's eyes); (*d*) the method is particularly costly with regard to materials; and, further, (*e*) the correction curve is based on all the data, whereas the working curve is based on only one-seventh of the total available data: from the fundamental assumption contained in equation (1) C_E and C_S are of equal importance and should therefore be given equal weight in all subsequent calculation. These considerations led me to re-examine the technique, which is otherwise sound, with a view to devising a quicker, less costly, but equally efficient method of calibration. As a result the following procedure was recommended.

Let the sets of standards of varying internal standard content be A, B, C, D, E, and F covering in equal steps the same C_S range as in the original method. From each set are selected standard dilutions corresponding to concentrations of 10, 31·6, 100 and 1000 p.p.m. of trace constituent. Single spectra are recorded for each standard of each set, giving in all twenty-four spectra, and the photometric evaluation is carried out in the usual manner to yield $\log (I_E/I_S)$ values. The working and correction curves are then obtained as follows:

Working Curve. From the available data a regression line of $\log (I_E/I_S)$ against $\log C_E$ is derived and used as the standard working curve. For convenience in notation let $\log C_E$ and $\log(I_E/I_S)$ be represented respectively by X and Y. The above regression line is made to pass through the point $(\bar{X}, \bar{Y}_C)$ where $\bar{X}$ = the mean log concentration = 1·875 and $\bar{Y}_C$ = the mean value of $\log(I_E/I_S)$ for the data from set C = $\overline{[\log(I_E/I_S)]}_C$. In general, for p sets of N points, the best estimate of the gradient of the plots of Y against X, assumed parallel through the p sets A, B, C, . . . , is given by

$$\bar{m} = \frac{\sum_A (X-\bar{X}_A)(Y-\bar{Y}_A) + \sum_B (X-\bar{X}_B)(Y-\bar{Y}_B) + \ldots}{\sum_A (X-\bar{X}_A)^2 + \sum_B (X-\bar{X}_B)^2 + \ldots}$$

Since here we have p identical sets of X values we can simplify this expression to

$$\bar{m} = \frac{\Sigma(X-\bar{X})(\bar{\bar{Y}}-G)}{\Sigma(X-\bar{X})^2}$$

where $\bar{\bar{Y}}$ = average value of Y over the p sets for a given value of X, and G = the grand average of Y for all 24 readings. The summation is over the N levels of the concentration.

Correction Curve. The mean value of $\log(I_E/I_S)$ is calculated for each set: A, B, etc. These values may be denoted $\overline{Y}_A$, $\overline{Y}_B$, etc. The general expression for the co-ordinates ($\%C_S$, correction) of the correction graph is then given by

$$\%C_{S(A,B,\ldots)}\ ,\quad \overline{Y}_C - \overline{Y}_{A,B,\ldots}$$

In subsequent estimations the correction factor is read from the correction graph and is added to the measured value of $\log(I_E/I_S)$; the value of $\log C_E$ is then obtained from the corrected value of $\log(I_E/I_S)$ and the working curve.

The advantages of this method are: (*a*) it is less time-consuming, only twenty-four as opposed to eighty-four spectra being required, with subsequent lessening in the time associated with operational technique, photometric evaluation, curve-drawing, etc.; (*b*) it is consequently less costly and especially so with regard to electrode material and photographic emulsion; (*c*) only eleven simple calculations are required from the data to construct the correction curve; (*d*) the best estimate of the working curve gradient is derived from all the available data; and (*e*) it is possible to determine limits of error for any concentration obtained throughout the range of trace-element concentration and over the range of internal standard content employed. The error involved in determining a concentration from an observed reading of $\log(I_E/I_S)$ is given by $\pm s_X t$ for a certain probability P for $(pN-p-1)$ degrees of freedom.

In the determination of the expression for s_X^2 (the variance of a value of X estimated by this method) it must be remembered that the use of a working curve of gradient $\bar{m}$ and correction factors obtained from the mean values is algebraically equivalent to the use of a series of parallel lines of gradient $\bar{m}$ through these mean values. It may be shown by standard statistical methods that a good approximation for s_X^2 is given by the following equation, provided that the standard error of $\bar{m}$ is small compared with $\bar{m}$, as it will be here.

$$s_X^2 = \frac{s_Y^2}{\bar{m}^2}\left\{1 + \frac{1}{N} + \frac{[\log(I_E/I_S) - \overline{\log\,(I_E/I_S)_C}]^2}{p\bar{m}^2\Sigma(X-\overline{X})^2}\right\}$$

where $s_Y^2 = \dfrac{\sum_A(Y-\overline{Y}_A)^2 + \sum_B(Y-\overline{Y}_B)^2 + \ldots - p\bar{m}^2\Sigma(X-\overline{X})^2}{pN-p-1}$

and the value of $\log(I_E/I_S)$ to be inserted in the above equation is the corrected value.

The above treatment applies strictly to the range 10-1000 p.p.m. To extend the working curve selected (C) to cover the 1000-3162 p.p.m. range it would only be necessary to take in addition duplicate spectra for the standard dilution (C, 3162 p.p.m.) this being necessitated by the falling-off of the curves; as the vertical displacements are reasonably constant over this range the procedure for the correction does not require any alteration.

The above equation for the variance of an estimated value of X applies strictly to any of the levels of C_S chosen for the sets of measurements used in the calibration. But in actual practice, further measurements will involve interpolation on a smoothed correction curve for other values of C_S. The smoothing would be expected to decrease the variance of the estimate, although the difficulty of drawing a freehand curve would probably make this decrease less than expected. Hence it seems reasonable to take the above value of s_X^2 as a rough upper bound for the variance of an estimate of X at a known value of C_S.

It should be noted that in the calibration C_S is accurately known: in subsequent routine analysis, however, this no longer holds and each sample must be separately analysed for internal standard content. The above limits of error $\pm s_X t$ make no allowance for the component of variance associated with the determination of C_S.

Operator-time and Cost Reduction. It is interesting to note the actual saving in time and cost introduced by the proposed scheme of calibration. It is found that one operator calibrating for six elements requires 18 hours for the task compared with 60 hours in the original scheme, thus effecting a saving of 70 per cent of the time normally expended in standardisation. Comparison of the costs for both methods (based on one operator/six-element calibration) shows a reduction in total operational cost (including time and materials) from £30 to £10. The corresponding reduction for materials alone (i.e. carbon electrodes and photographic emulsion) is from £13 to £3. If an annual calibration is considered sufficient, and this decision must be left to the analyst in charge, then a total saving of £20 is effected. In the small type of instrumental laboratory which I have in mind this can represent an economy of expenditure of the order of 10 per cent. Alternatively, for reasons which will only be apparent to the analyst (one might be the possibility of 'instrumental fatigue' through constant usage), it might be advantageous to check the instrumental performance thrice annually and this could be done without exceeding the limits of expenditure associated with the original method of calibration.

Photometric Measurement

In flame-spectrographic methods[3] it is usual to plot θ_B/θ_E against concentration,

where θ_E = galvanometer deflection for a particular spectral line of the element;

θ_B = corresponding deflection for the background (e.g. the spectrum of the carbon monoxide flame from an air-acetylene source) measured at some convenient specified point in the neighbourhood of the spectral line concerned.

This background varies in density throughout the length of the spec-

trum, but at any particular point (i.e. wavelength) it should be reproducible for a series of spectra. Its introduction in the method therefore compensates for fluctuations in the light source or small variations in total emission from one spectrum to the next. In other words it serves as a measure of flame constancy or as an internal standard. Each photographic plate carries, in addition to the sample spectra, a number of standard spectra from which the θ_B/θ_E versus C_E relationship is derived. In determining the elements X, Y, it is necessary to obtain four readings—$\theta_B(X)$, $\theta_E(X)$, $\theta_B(Y)$, $\theta_E(Y)$. When, however, the spectral lines for X and Y lie very close in the spectrum, the question arises: can one background reading serve for both lines? For example, in the determination of potassium (K) and manganese (Mn) by this method using the K, 4044 Å and Mn, 4031 Å lines, I obtained the background readings shown in the following table. These were obtained from a plate carrying thirty spectra.

TABLE I

Background reading, θ_B (cm)

θ_B(K)	θ_B(Mn)	θ_B(K)	θ_B(Mn)	θ_B(K)	θ_B(Mn)
26·5	26·8	24·8	24·7	25·9	26·3
26·4	26·9	24·7	24·9	26·5	27·3
27·2	26·6	25·8	25·9	26·5	26·4
27·2	27·4	26·2	26·0	26·8	26·7
27·2	27·3	26·9	25·8	26·6	26·5
26·6	25·9	26·2	26·1	27·0	27·3
27·4	26·0	26·0	25·8	26·2	27·7
27·3	27·3	26·3	26·7	26·2	26·2
26·8	26·4	26·6	26·2	26·4	26·7
26·4	26·1	26·0	25·8	26·1	26·8

The differences are small but not consistent; if we regard them as a sample from the hypothetical population formed by the differences from an infinite number of future assays, it may be determined whether the observed values could in fact have been drawn from a population where the mean difference is zero. With the usual notation and writing the differences in background as X, we have:

$$\begin{array}{lll} N = 30 & \Sigma(X^2) = 8{\cdot}92 & s^2 = 0{\cdot}307\,54 \\ \Sigma(X) = 0{\cdot}2 & (\Sigma X)^2/N = 0{\cdot}001\,33 & s = 0{\cdot}554 \\ \bar{X} = 0{\cdot}0066 & \Sigma(X-\bar{X})^2 = 8{\cdot}918\,67 & s_{\bar{X}} = 0{\cdot}1011 \end{array}$$

The statistic t is calculated from

$$t = (\bar{X} - \mu)/s_{\bar{X}}$$

where μ is the hypothetical population mean.
On the null hypothesis, $\mu = 0$
and

$$t = (0{\cdot}0066 - 0)/0{\cdot}1011 = 0{\cdot}065$$

Entering the t table with d.f. $= 29$, we have $t = 2{\cdot}045$ at the 5% level. The mean is therefore not significantly different from zero. Further tests confirmed these results and on the basis of these the null hypo-

thesis was accepted. Hence there is no need to read both backgrounds and in this case the background for the manganese line was recommended for use from the point of view of operational suitability. The above modification for the flame-spectrographic determination of K and Mn in solution further demonstrates the saving in operational time which can be effected by correct interpretation of the data.

Conclusions

In the examples discussed it is obvious that much operational time can be saved and unnecessary work eliminated in the instrumental laboratory. One further point may be mentioned. Both instrument manufacturer and analyst will endorse the statement that one of the main advantages of instrumental methods lies in their speed; it is therefore essential that this gain over the more conventional analytical methods should not be wholly or even partially sacrificed through improper presentation and evaluation of the analytical data.

REFERENCES

[1] Scott, R. O. (1946). *J. Soc. Chem. Ind., Lond.*, **65**, 291.
[2] Mitchell, R. L. (1948). *The Spectrographic Analysis of Soils, Plants and Related Materials*. Commonwealth Agricultural Bureau Tech. Comm. No. 44.
[3] Black, I. A. and Smith, E. (1950). *J. Sci. Fd. Agric.*, **1**, 201.

PHYSICAL INVENTORY USING SAMPLING METHODS

MARION R. BRYSON
Duke University, Durham, North Carolina

An integral part of any commercial activity is the periodic counting of assets. In some larger establishments this is a major undertaking involving many man-hours of labour. In this paper Dr Bryson presents a method of continuous inventory in which some quality control methods are used. The basic idea is to use sampling to determine when and where an inventory should be taken rather than use the customary method of taking an annual inventory at a pre-set time. The results of the application of this method at two large government supply centres are described, and the method is compared with the annual inventory method.

Introduction

Business, industry, and government are all faced with the task of periodically taking a physical inventory of all goods on hand. The most commonly used method of performing this physical inventory is to make a complete count of all stocks once each period, usually annually. All counts of stocks are made within a short period of time, say within a week. At the close of such an inventory, it is assumed that accurate knowledge of the quantity and value of all stocks on hand has been obtained. Unfortunately, in most cases this may be a false assumption.

In most of the larger and many of the smaller establishments a continuous record is kept of stocks. These records are altered as activity changes the quantity on hand. If the records were completely accurate it would, obviously, be unnecessary ever to make a physical count of stocks. Many of the records are inaccurate for the following reasons: the previous inventory was in error causing the record to reflect the wrong balance from the beginning of the fiscal period; some receipts and/or issues were made without correct alteration of the records; items were lost; items were pilfered; records were lost; and other record adjustments, such as price changes or reserve stock levels, were incorrectly processed. Hence it is the purpose of the inventory to correct the records which reflect wrong information.

The question which immediately arises is 'How accurate are the records at the close of the inventory?' The personnel employed in the inventory are relatively untrained in inventory methods since this activity occurs during only one week out of the year. During this week personnel are drawn from all parts of the establishment to perform the counting and clerical duties incidental to the inventory. Many of these people are not interested in inventory work, and if they are

doing satisfactory work at their normal job they know they will not be dismissed or demoted if their poor inventory work is detected.

Because of the suspected inaccuracies in inventory and because operational record errors occur and remain undetected for as long as one year, an agency of the US Government sponsored a research project to determine what could be done about improving inventory methods.

Description of Supply Centres

The government agency mentioned above maintains supply centres in various parts of the country. Each of these supply centres has in its storage areas from 5 000 to 150 000 different types of item stored. Hereafter the term 'item' will be used to denote a type of item and the term 'piece' to denote an individual part. From zero to more than one million pieces of each item may be on hand at any given time. The storage areas are in the nature of huge stockrooms whose function it is to receive and store incoming goods and issue and ship goods ordered by their customers. The supply centre maintains records of stocks on hand and the records department receives orders for shipments, processes them, and instructs the stockroom to issue the stock. It also processes all receipt vouchers and instructs the stockroom to store the stock in an assigned location.

Each item has an assigned location or locations in the stockrooms. When an order comes in to the records department, the records are checked to determine whether the stock in this quantity is available for issue. If it is not, the order is sent to a master records centre which will re-order the stock from another storage centre. If the stock is available, a shipping order is drawn up and sent to the stockroom. This shipping order designates the item number, location, and quantity to be shipped, among other facts. If the item is stored in more than one location, one of the locations is designated as the master location. The recorded balances are kept by item number and not by location; hence the single balance kept by the records department reflects the sum of the quantities in each storage location. The records department adjusts its balance to reflect the issue. Receipts are processed in essentially the same way.

Basic Plan of Operation

One such supply centre was selected by the government agency as the site of the study. After nine months of preliminary work at this centre, an additional centre was assigned for further study and testing. Each of these two centres has approximately 100 000 items in its stockrooms.

Various methods of quality control have been used from time to time in the actual taking of a shut-down inventory. These have been principally confined to quality control of the individual counter and quality control of the areas of the stockroom after the counting was

complete. These methods have not proved to be very successful in the inventories taken by the agency which sponsored this research.

The basic concept of the present project is the continuous quality control of the storage areas. Those areas which are deemed to be out of control are completely inventoried. This serves the same purpose as 100% inspection of a rejected lot of pieces produced on a production line.

The stockrooms of a supply centre are stratified by physical location. All lots are of approximately equal size. An attempt is made to store the more active items together in one or more lots and to store the less active items in different lots. In addition to the advantage this 'popularity' storage has of placing the active items nearer the shipping facilities, it also serves to increase the between-lot variance of the error rates.

Quality Control Procedure

To initiate the quality control once the stratification is complete, a sample of items in each lot is selected. A physical count of the number of pieces of each of these items in the stockroom is made and this count is compared with the recorded balance. If the count disagrees with the balance by more than the 'leeway', the item is said to be 'discrepant' or 'in error'. This leeway is established for the discrepancies so that a minor discrepancy may be ignored. In this study an item is classed as discrepant if and only if the difference between the count and the balance is greater than 1% of the balance, or the monetary value of the discrepancy is greater than $1 (approximately 7s.).

Following the count, an estimate of the percentage of items in each lot which are discrepant is made. The simple binomial estimator

$$\hat{P}_i = \frac{100 d_i}{n_i}$$

is used. Here

$\hat{P}_i$ = the estimated percent discrepant in the i-th lot; (hereafter called 'estimated error rate').
d_i = number of discrepant items in the sample from the i-th lot.
n_i = sample size in the i-th lot, i.e. the number of items.

The variance of this estimator is the usual one:

$$\text{var } \hat{P}_i = \frac{\hat{P}_i(100 - \hat{P}_i)}{n_i}$$

Next, an acceptable limit for the estimated error rate is established. In this project the limit is 10%.

Hereafter the words 'sampling' or 'sample' refer to the process of taking a physical count on only a sample of items in a lot. The word

'inventory' will refer to a complete count of every item in the lot and the reconciliation of this count with its recorded balance.

The results of the sample are then observed. If more lots have an estimated error rate of above 10% than can be inventoried by the counting crew in a three-month period (quarter), only enough of the highest error-rate lots are rejected to consume one quarter of inventory time. The remainder of those above 10% are temporarily accepted.

In the first quarter of the year, those lots rejected on the initial sample are inventoried. Those lots which were accepted, whether or not their estimated error rates were above 10%, are resampled as before during this quarter. From this group of resampled lots the inventory load for the second quarter is chosen. It will be seen that those lots inventoried in the first quarter cannot be inventoried again during the second quarter since they were not included in the re-sampled lots during the first quarter. They will be resampled during the quarter following their inventory, i.e. during the second quarter in this case, so cannot be rescheduled for inventory before the third quarter.

If the lots whose estimated error rates are above 10% do not constitute a full quarter of inventory work, either the acceptable limit can be lowered or the size of the inventory crew can be reduced.

Operation of Initial Study

In October of 1958 the project was initiated at supply centre number one (SC-1). For the first nine months the work was confined to methods of stratification and methods of counting an item without freezing activity on the item. A team of 15 men was assigned to work on the project. The centre was stratified into 25 lots and two samples were taken in each lot, the time between samples being about four months.

A completely random sample of items in each lot was drawn. First, for sampling purposes, every piece of a given item was considered as being stored in the lot with the master location of the item. A cumulative list of all possible storage locations was drawn up. This list is not the same as that kept in the 'location file' from which issues are drawn up. For example, a given rack may have as many as three items stored on it, all of which have the same numbered location in the location file. This rack was considered as three 'item locations' even though the records showed it as only one location. It was because of this that it was not possible to sample from the location file. This would result in clusters of items in the sample, the cluster size varying from one to twenty-four items.

The sample size was not rigidly fixed in advance. A sample of item locations was drawn from the cumulative list, using random number tables, and each of these locations was visited. If the location was the master location of some item, this item was included in the sample. If it was a reserve location or an empty location to which no stock

number was assigned, the location was dropped from the sample. If the final sample size was too small, a second sample was taken to augment the first one. A sample size of between 200 and 300 was selected in all but a few small lots. The average lot size was about 4000.

The count on the sample items was taken without affecting the normal centre operation. In simplified form the steps followed were:

(1) Count all pieces of an item in the stockroom.
(2) Obtain the recorded balance for the item.
(3) Compare the two figures.
(4) If the figures agree, *classify it as a non-discrepant item.*
(5) If the figures disagree, they may disagree because
  (*a*) A miscount has occurred.
  (*b*) An activity has occurred in the record section but has not yet occurred in the stockroom.
  (*c*) The item is discrepant.
  This item is set aside for two weeks.
(6) Two weeks after the first count, retrace steps 1-4 if the item is in disagreement.
(7) If the two figures now agree, *it is a non-discrepant item.*
(8) If they disagree by the same quantity as they did on the first count, *the item is classed as a discrepant item.* Points (*a*) and (*b*) in step 5 are improbable in this case since the same miscount is unlikely and no activity should be in process for two weeks.
(9) If the item disagrees by an amount different from the first discrepancy, *set it aside for another two weeks.*
(10) A third count is taken on all items falling into step 9 and steps 1-4 and 8 are followed.
(11) For items which are still not reconciled, an investigation is conducted to determine their true nature.

Several minor refinements can be and have been made in the above procedure but basically it has been operated successfully using this system.

The purpose of the second sampling was to estimate the error rates which existed in each of the 25 lots. The error rates obtained in this sample were then compared with the results of the annual shut-down inventory which was taken immediately following the second sampling phase.

Results of the Initial Study

Table I gives the lot sizes, sample sizes, and the error rates of both the sample and the shut-down inventory.

From the table it will be noted that the difference between the sample results and the findings of the inventory is significant at the 95% level for more than half of the lots. Since it was the purpose of both the sample and the inventory to estimate the true error rate of

TABLE I

Error rates (%)

Lot No.	Lot Size	Sample Size	Estimated Error Rate from Sample	Shut-down Inventory Error Rate	Difference (Inv. – Samp.)
1	7 818	310	61·9	52·3	–9·6†
2	7 191	291	57·7	53·0	–4·7
3	1 809	217	51·6	54·8	3·2
4	2 344	66	46·9	50·9	4·0
5	6 717	285	43·5	50·5	7·0*
6	3 228	220	41·8	32·9	–8·9*
7	7 945	300	39·0	41·5	2·5
8	3 493	212	37·3	47·0	9·7*
9	3 676	207	37·2	31·2	–6·0
10	3 136	217	36·9	53·7	16·8†
11	1 908	207	35·3	33·3	–2·0
12	5 053	259	34·8	50·5	15·7†
13	7 271	287	33·1	47·5	14·4†
14	2 930	213	31·9	34·3	2·4
15	6 320	239	31·4	37·5	6·1*
16	2 885	234	29·5	41·0	11·5†
17	614	72	29·2	39·9	10·7*
18	3 084	222	27·5	37·4	9·9†
19	7 074	233	27·0	25·7	–1·3
20	1 918	226	26·5	31·2	4·7
21	2 266	197	24·4	28·6	4·2
22	3 175	216	22·2	33·4	11·2†
23	6 248	260	21·9	25·4	3·5
24	1 401	192	20·3	25·5	5·2
25	2 872	242	16·5	25·6	9·1†
TOTAL	102 376	5 624	—	—	—
MEAN	—	—	34·7	40·7	6·0

* Difference significant at 95% confidence level
† Difference significant at 99% confidence level

the lot it appears that something is wrong. It could be one or both of two things:

(1) the sample was drawn incorrectly, biasing the results;
(2) the sample does not estimate the percentage found by the inventory but some other quantity; either or both may differ from the true error rate.

Let us look at the first of these two possibilities. In order to check the accuracy of the sampling, the *inventory* findings on the *sample items only* were compared with the inventory findings on the entire lot. The difference between these two figures represents sampling error only. Table II gives the results of this comparison.

In this comparison only three of the twenty-five lots had differences which were significant at the 95% level. This leads one to believe that the sampling was without bias with the possible exception of lots 1 and 10. Since one of these is negative and the other is positive, and

since there are equal numbers of plus and minus differences, no consistent bias is evident.

Let us look now at the second possible cause of the large differences shown in Table I.

Since the differences in error rates given by Inventory (Inventory denotes this particular inventory) and the estimates given by the sampling plan (hereafter referred to as SQC) are not due primarily to the sampling technique of SQC, these differences must be due in large part to mechanical mistakes by SQC, or mistakes by Inventory, or change in discrepancy status of items from one count to the other. The latter group is regarded as not significant. It becomes then a question of determining the relative order of magnitude of the mistakes made by SQC and those made by Inventory.

TABLE II

Error rates (%)

Lot No.	Inventory Error Rate on Sample Items	Inventory Error Rate on All Items in Lot	Difference (Lot – Sample)
1	61·3	52·3	–9·0†
2	55·5	53·0	–2·5
3	57·8	54·8	–3·0
4	47·8	50·9	3·1
5	50·7	50·5	–0·2
6	39·3	32·9	–6·4
7	46·2	41·5	–4·7
8	45·5	47·0	1·5
9	41·4	31·2	–10·2*
10	43·3	53·7	10·4†
11	39·2	33·3	–6·2
12	46·5	50·5	3·9
13	43·8	47·5	3·7
14	32·7	34·3	1·6
15	35·5	37·5	2·0
16	40·6	41·0	0·4
17	40·6	41·0	0·4
18	34·7	37·4	2·7
19	29·3	25·7	–3·6
20	31·3	31·2	–0·1
21	28·6	28·6	0·0
22	29·2	33·4	4·2
23	30·4	25·4	–5·0
24	26·8	25·5	–1·3
25	22·4	25·6	3·2
MEAN	39·9	40·7	0·8

* Difference significant at 95% confidence level
† Difference significant at 99% confidence level

Information on the mistakes of SQC and those of Inventory is provided through a study of the items counted by both SQC and Inventory. The results of this study follow.

1. Inventory and SQC agree on a total of 81 per cent of the items, the non-discrepant items in agreement being 58 per cent, and the discrepant items 23 per cent.

2. Inventory found a discrepancy, SQC found no discrepancy, on 7·75 per cent of the items. In a further study it was found that Inventory mistakes account for 7 per cent of these, and the other 0·75 per cent are actual discrepancies which occurred after the sampling count was completed.

3. SQC found a discrepancy, Inventory found no discrepancy, on approximately 1·5 per cent of the items. Essentially all of these differences result from SQC mistakes.

4. SQC and Inventory both found a discrepancy, but disagreed about its magnitude, on 9·75 per cent of the items. By special investigation on these items, it was found that SQC and Inventory shared about equally in the mistakes, but that about 2% of the mistakes were due to an actual change in the size of the discrepancy between the time of the sample count and the time of the inventory count.

From this analysis we arrive at the figures in Table III. In accordance with this table, for example, the SQC technique results in 22 mistakes for every 1000 good items counted, and in 122 mistakes for every 1000 discrepant items counted. The rate of mistakes in general is 5·5 per cent of the time for SQC and 11 per cent of the time for Inventory. The general conclusion is that overall the SQC technique makes only half as many mistakes as the Inventory technique.

TABLE III

Frequency of mistakes, Inventory compared with SQC

	Non-discrepant Items	Discrepant Items	All Items
SQC	2·2%	12·2%	5·5%
Inventory	10·4%	12·2%	11·0%

This explains the differences shown in Table I. Of the thirteen significant differences shown there, only two of them are negative and in these two cases the difference is largely explained by sampling bias or sampling error, as indicated in Table II.

There are two major reasons for the superior performance of the SQC team. They are:

(1) The procedure of accepting no count unless it agrees with a previous count or with the recorded balance virtually eliminates the possibility of a miscount.
(2) The team is experienced in inventory and interested in the work.

Results of Second Study

After the first study at SC-1, the project was initiated at a second supply centre (SC-2). It was also expanded to include the physical inventory of the rejected lots. In these two centres, no shut-down inventory is being taken.

After some study it was decided that a centre should be divided into sixteen lots of equal size. If more than three lots had an estimated error rate of more than 10%, the three lots showing the highest percentage error in any quarter would be completely inventoried the following quarter. The principal advantage of this is that the work load in each quarter is the same so that it is not then necessary to have a variable-sized inventory team.

TABLE IV

Comparison of rankings of error rates of lots by different methods

Rank by Single Count	Rank by Multiple Count	Rank by Inventory
1	1	4
2	2	3
3	3	1
4	4	5
5	5	6
6	6	18
7	7	10
8	8	9
9	11	17
10	9	19
11	10	2
12	12	7
13	17	12
14	16	11
15	13	8
16	14	15
17	15	13
18	18	14
19	20	20
20	19	12
21	21	21
22	23	25
23	22	16
24	24	24
25	25	23

In any given quarter thirteen lots are sampled and three lots are inventoried so that each year an equivalent of three-quarters of the lots in the entire centre is inventoried completely. Actually some lots may be inventoried twice in a year so that more than one-quarter of the centre may miss inventory entirely. Adjustments in the recorded balances for discrepant items are made.

It might at first sight be thought paradoxical that the recommended plan should include a proportion of complete inventory work, since the earlier results indicated that more mistakes were made in a complete inventory than in a sample count. However, it must be stressed that the earlier comparison was between an experienced SQC team and relatively untrained personnel carrying out the inventory. A team specialising in inventory work would be expected to have a low mistake rate. Furthermore, the periodic complete inventory allows proper adjustment of erroneous balances to be made.

Since the purpose of the sampling is to rank the lots from most erroneous to least erroneous, the question whether recounts were necessary in the sampling arose. If only one count is taken on the sample items and the item is classed as discrepant if its count does not agree with its balance, ignoring the possibility of miscounts and documents in flow, what effect would this have on the ranking? This question was studied using the data on the 5624 items in the sample at SC-1. Each item was classed as a discrepant or a non-discrepant item on the basis of the first count only and the lots were ranked. The ranking thus obtained was compared with the ranking using the recount procedure. The results of this are presented in Table IV.

Spearman's rank correlation coefficient has the following values:

(1) single count *v.* multiple count $r_s = 0{\cdot}984$
(2) single count *v.* inventory $r_s = 0{\cdot}754$
(3) multiple count *v.* inventory $r_s = 0{\cdot}775$

The single count agrees extremely well with the multiple count and both are in reasonable agreement with the inventory findings.

On the basis of the foregoing results, it was decided that for the sample items, only a single count would be taken. This saves considerable time in the counting. No adjustments are made on the basis of the sample findings.

During the first quarter, SC-2 was sampled, without stratification. On the basis of this sample, the error rate in each section of each stockroom was estimated. The sections were then grouped together into lots, sections with similar estimated error rates being put into the same lots. The three lots with the highest error rates were inventoried, and the remaining lots were resampled in the following quarter. In

TABLE V

Results of study at SC-2

Lot No.	Error Rate on First Sample (%)	Error Rate on Inventory (%)	Error Rate on Second Sample (%)
1	40·0	29·5	5·9
2	39·8	32·6	5·5
3	34·9	30·1	4·9
4	26·7		24·4
5	25·6		30·1
6	24·0		18·8
7	20·6		24·2
8	19·5		30·1
9	19·4		12·0
10	14·3		11·5
11	12·7		18·8
12	12·0		13·2
13	11·8		10·0
14	11·7		10·4
15	11·1		14·3
16	*		*
Mean	21·6	30·7	15·6
Mean of Lots 4-15	17·5		18·2

* No results available

the quarter following the inventory, the three inventoried lots were resampled.

The results of the samples and the inventories at SC-2 are given in Table V.

It will be noted that the inventory in the first three lots reduced the error rate radically. It had an overall effect of reducing the centre-wide error rate by six percentage points. It will be noted also that the error rate in the lots which were not inventoried did not grow significantly during the quarter.

There are no significant differences between the sample estimates in the first and second quarters except for lot 8. In this instance, much stock was moved into the lot during the quarter. The stock which was moved in was believed to have a high error rate by the centre personnel.

Comparison of Inventories

The direct cost per item of the SQC inventory is about 70-80% of the direct cost of the shut-down inventory. The saving in cost is the result of more efficient operation of the SQC inventory and the lack of necessity for training a large group of people for inventory. Other advantages of the SQC inventory are

(1) No shut-down period is necessary, so that normal operation is carried on throughout the year.
(2) At any given time the records have a higher degree of accuracy.
(3) A trained team is available for any special inventories which may be necessary.
(4) Inefficient or error-making procedures are quickly identified and eliminated.
(5) An incentive for more efficient centre operation is produced.
(6) State of preservation of stocks is constantly reviewed.

Acknowledgement

I wish to express my thanks to Professor R. F. Rinehart, who critically read the manuscript and made many helpful suggestions.

ANALYSIS OF VARIANCE OF A BALANCED INCOMPLETE BLOCK DESIGN WITH MISSING OBSERVATIONS*

HUGH ROBERT BAIRD and CLYDE YOUNG KRAMER
Virginia Agricultural Experiment Station of the Virginia Polytechnic Institute

The authors obtain explicit formulae for all cases when two values are missing, formulae for some other specific cases, and a general solution. They give a direct method of analysing the augmented data and an illustrative example.

Introduction

In experimental work the results of one or more observations may be missing. An agricultural plot may be trampled, an animal may die, a test tube may be dropped, or a machine may break down. When missing values occur, the usual method of computing the various sums of squares cannot be used unless the missing values are first estimated from the existing data.

Cornish[1] obtained a specific formula for a single missing value and an iterative procedure for several missing values by minimising the error sum of squares. He also showed that the resulting treatment sum of squares had a positive bias and how to eliminate it. Wilkinson[3] obtained a general procedure for estimating missing values which involves the solutions of matrix equations. The method of Glenn and Kramer[2] may also be used by considering that a balanced incomplete block design is a randomised block design with missing observations, but this procedure would require the estimation of a large number of missing values.

This paper will deal with the estimation of several missing values in a balanced incomplete block design by minimising the error sum of squares. Explicit formulae for each value missing will be derived for many special cases and a completely general solution will be given. These formulae may prove to be less tedious in application than the ones now available. A direct method of analysis of the augmented data, not requiring a correction for bias in the treatment sum of squares, will be given.

It should be noted that, by estimating missing values, the symmetry in the design is restored, but the estimates serve only to expedite analysis of the data; they do not by any means restore the information lost in the missing observations.

* A condensation of a thesis by H. R. Baird written under the direction of C. Y. Kramer and submitted to the Virginia Polytechnic Institute in partial fulfilment of the requirements for the M.S. degree in Statistics.

Research sponsored in part by Division of Research Grants, Contract No. 2G and National Institute of Mental Health, United States Public Health Service.

Model and Notation

The experimental model for this design is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}; \; i = 1, \ldots, t; j = 1, \ldots, b \quad \ldots . (2.1)$$

where μ is the overall mean, τ_i is the effect of the ith treatment, β_j is the effect of the jth block, and ϵ_{ij} is a random effect assumed to be normally distributed with mean 0 and variance σ^2; the variates ϵ_{ij} are further taken to be stochastically independent. It should be noted that only certain combinations of values of i and j are possible and these are determined by the nature of the experimental design.

We shall denote treatment totals and block totals by T_i and B_j if there are no missing values in the ith treatment or jth block and let T_i' and B_j' equal the total of observed values in the ith treatment and jth block if there are missing values. We also define

$$Q_i = kT_i - B_{\cdot i} \quad \ldots . (2.2)$$

where k is the block size and $B_{\cdot i}$ is the sum of the B_j's for those blocks containing treatment i.

When a particular missing value is being considered, this missing value will be referred to as x_{cd}. Two particular missing values will be referred to as x_{cd} and x_{ef}; n missing values will be denoted by x_{cd}, $x_{ef}, \ldots, x_{uv}$.

Since x_{cd} will be involved in two types of Q-values, the Q-value for treatment c and Q-values for the other treatments in the same block as x_{cd}, we denote the former by

$$Q_c = Q_c' + (\text{terms involving missing values}) \quad \ldots . (2.3)$$

and the latter by

$$Q_i = Q_{i(c)}' + (\text{terms involving missing values}) \quad \ldots . (2.4)$$

where Q_c' and $Q_{i(c)}'$ are the numerical values obtained from observed data, and defined by an equation of the same form as (2.2). The necessary calculations are explained in detail in the illustrative example.

If we differentiate the error sum of squares with respect to a particular missing observation, say x_{cd}, and equate this partial derivative to zero, we can express the equation so that the left-hand side contains only x's and the right-hand side contains only B_d', Q_c', and $\Sigma Q_{i(c)}'$. Now the form of the right-hand side of the equation will always be the same, except for a change of subscripts, regardless of the missing value with respect to which the error term is differentiated. We therefore denote, for expediency, the right-hand side of the equation obtained by differentiating the error sum of squares with respect to the missing observation, x_{cd}, as C_{cd} where

$$C_{cd} = rt(k-1)B_d' + (t-1)(k-1)Q_c' - (t-1)\sum_{i \neq c} Q_{i(c)}' \quad \ldots . (2.5)$$

Matrix Representation and Solution

If there are n missing values, $x_{cd}, x_{ef}, \ldots, x_{uv}$, and if the error sum of squares is differentiated with respect to each of the missing values in the order specified, the result, upon equating these partial derivatives to zero, is the equation system

$$\begin{aligned} mx_{cd} &+ q_{12}x_{ef} + \cdots + q_{1n}x_{uv} = C_{cd} \\ q_{21}x_{cd} &+ mx_{ef} + \cdots + q_{2n}x_{uv} = C_{ef} \\ &\cdots \\ q_{n1}x_{cd} &+ q_{n2}x_{ef} + \cdots + mx_{uv} = C_{uv} \end{aligned} \qquad \ldots . \quad (3.1)$$

where

$$m = (k-1)(rtk - rt - tk + k) \qquad \ldots . \quad (3.2)$$

and

$$q_{12} = q_{21},\ q_{13} = q_{31}, \text{ etc.}$$

The q's are obtained by the appropriate combination of coefficients from Table I, which have been obtained by considering all possible configurations of the pair of missing values, x_{cd} and x_{ef}, differentiating always with respect to x_{cd}. It is extremely important to observe that these configurations are not all mutually exclusive. (With n missing values, all possible pairs must be considered. Table I applies equally well to any other pairs such as x_{gh} and x_{uv}, with suitable change of subscripts in the Table.)

For instance, x_{cd} and x_{ef} may occur in the same treatment; in addition, there may be three other treatments replicated both in block d and in block f. Thus x_{cd} and x_{ef} would fit both configuration 3 and configuration 4. The coefficients of x_{ef} would be added and the appropriate q value would be

$$q_{12} = (t-1)[3 + (k-1)^2] \qquad \ldots . \quad (3.3)$$

The equation system (3.1) may be written in matrix form as

$$\mathbf{Ax} = \mathbf{C} \qquad \ldots . \quad (3.4)$$

TABLE I

*Off-diagonal coefficients**

Configuration of x_{cd} and x_{ef}	Coefficient of x_{ef}
1. No treatments common to blocks d and f.	0
2. x_{cd} and x_{ef} occur in the same block.	$-\frac{m}{k-1}$
3. x_{cd} and x_{ef} occur in the same treatment; no other treatments common to both blocks.	$-(t-1)(k-1)^2$
4. There are p treatments (not c or e) common to blocks d and f.	$-p(t-1)$
5. Treatment c is replicated in block f OR treatment e is replicated in block d; no other treatments common to both blocks.	$(t-1)(k-1)$
6. Treatment c is replicated in block f AND treatment e is replicated in block d; no other treatments common to both blocks.	$2(t-1)(k-1)$

* The Table can be used for any other pairs, such as x_{gh} and x_{uv}, by changing the subscripts in the body of the Table.

where $\mathbf{A} = \begin{bmatrix} m & q_{12} & \cdots & q_{1n} \\ q_{21} & m & \cdots & q_{2n} \\ \cdot\cdot & & \cdot & \cdot \\ q_{n1} & q_{n2} & \cdots & m \end{bmatrix}$ .... (3.5)

$$\mathbf{x} = \begin{bmatrix} x_{cd} \\ x_{ef} \\ \cdot \\ \cdot \\ \cdot \\ \\ x_{uv} \end{bmatrix} \quad \ldots\ldots (3.6)$$

and

$$\mathbf{C} = \begin{bmatrix} C_{cd} \\ C_{ef} \\ \cdot \\ \cdot \\ \cdot \\ \\ C_{uv} \end{bmatrix} \quad \ldots\ldots (3.7)$$

The estimates of the missing values are then obtained from

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{C} \quad \ldots\ldots (3.8)$$

Case of n Missing Values in n Distinct Blocks and n Distinct Treatments; No Treatment Replicated in Any Two of the Involved Blocks

Clearly, every possible pair of missing observations satisfies configuration 1, so that (3.1) reduces to

$$\begin{array}{llll} mx_{cd} + 0 & + \ldots + 0 & = C_{cd} & \\ 0 + mx_{ef} & + \ldots + 0 & = C_{ef} & \ldots\ldots (4.1) \\ \cdot \quad\quad \cdot & \quad\quad \cdot & & \\ 0 \quad + 0 & + \ldots + mx_{uv} & = C_{uv} & \end{array}$$

In this case, the estimate of the missing value in treatment i and block j is simply

$$x_{ij} = \frac{rt(k-1)B'_j + (t-1)(k-1)Q'_i - (t-1)\sum\limits_{i' \neq i} Q'_{i'(i)}}{(k-1)(rtk - rt - tk + k)} \quad \ldots\ldots (4.2)$$

which is identical to the estimate of a single missing value as first shown by Cornish.

Case of n Missing Values in the Same Block

Here, each pair of missing values satisfies configuration 2, leading to the equation system

$$\begin{aligned} mx_{cd} &- \frac{m}{k-1}x_{ed} - \ldots - \frac{m}{k-1}x_{ud} = C_{cd} \\ -\frac{m}{k-1}x_{cd} &+ mx_{ed} - \ldots - \frac{m}{k-1}x_{ud} = C_{ed} \\ \cdot \quad & \cdot \qquad\qquad \cdot \\ -\frac{m}{k-1}x_{cd} &- \frac{m}{k-1}x_{ed} - \ldots + mx_{ud} = C_{ud} \end{aligned} \qquad \ldots . \quad (5.1)$$

Performing the operation, $\mathbf{A}^{-1}\mathbf{C}$, one obtains the missing value estimates

$$x_{id} = \frac{1}{mk(k-n)}\{[k(k-n)+(n-1)]C_{id} + (k-1)\sum_{i' \neq i} C_{i'd}\} \qquad \ldots . \quad (5.2)$$

Case of n Missing Values in the Same Treatment; No Other Treatment Replicated in Any Two of the Involved Blocks

In this case, every pair of missing values satisfies configuration 3, and the equation system is

$$\begin{aligned} mx_{cd} - (t-1)(k-1)^2x_{cf} - \ldots - (t-1)(k-1)^2x_{cv} &= C_{cd} \\ -(t-1)(k-1)^2x_{cd} + mx_{cf} - \ldots - (t-1)(k-1)^2x_{cv} &= C_{cf} \\ \cdot \qquad\qquad \cdot \qquad\qquad \cdot & \\ -(t-1)(k-1)^2x_{cd} - (t-1)(k-1)^2x_{cf} - \ldots + mx_{cv} &= C_{cv} \end{aligned} \qquad \ldots . \quad (6.1)$$

The solution yields the missing value estimates

$$x_{cj} = \frac{1}{K}\{[m-(n-2)(t-1)(k-1)^2]C_{cj} + (t-1)(k-1)^2\sum_{j' \neq j} C_{cj'}\} \qquad \ldots . \quad (6.2)$$

where $K = [m^2 - m(n-2)(t-1)(k-1)^2 - (n-1)(t-1)^2(k-1)^4]$ $\ldots . \quad (6.3)$

Case of Two Missing Values

In this section we shall first identify the configuration of the two missing values, x_{cd} and x_{ef}, and then give the explicit formulae for the estimates.

1. No treatments common to blocks d and f. Then

$$\begin{aligned} x_{cd} &= C_{cd}/m \\ x_{ef} &= C_{ef}/m \end{aligned} \qquad \ldots . \quad (7.1)$$

2. There are p treatments (not c or e) common to blocks d and f. Then

$$\begin{aligned} x_{cd} &= \frac{1}{[m^2 - p^2(t-1)^2]}\{mC_{cd} + p(t-1)C_{ef}\} \\ x_{ef} &= \frac{1}{[m^2 - p^2(t-1)^2]}\{mC_{ef} + p(t-1)C_{cd}\} \end{aligned} \qquad \ldots . \quad (7.2)$$

3. There are p treatments (not c or e) common to blocks d and f. It is possible that $p=0$. In addition, treatment c is replicated in block f OR treatment e is replicated in block d. In this event

$$x_{cd} = \frac{1}{[m^2-(t-1)^2(k-1-p)^2]}\{mC_{cd}-(t-1)(k-1-p)C_{ef}\} \quad \ldots . \quad (7.3)$$

$$x_{ef} = \frac{1}{[m^2-(t-1)^2\,(k-1-p)^2]}\{mC_{ef}-(t-1)(k-1-p)C_{cd}\}$$

4. There are p treatments (not c or e) common to blocks d and f. It is possible that $p=0$. In addition, treatment c is replicated in block f AND treatment e is replicated in block d. Then

$$x_{cd} = \frac{1}{[m^2-(t-1)^2\,(2k-2-p)^2]}\{mC_{cd}-(t-1)(2k-2-p)C_{ef}\} \quad \ldots . \quad (7.4)$$

$$x_{ef} = \frac{1}{[m^2-(t-1)^2\,(2k-2-p)^2]}\{mC_{ef}-(t-1)(2k-2-p)C_{cd}\}$$

5. x_{cd} and x_{ef} occur in the same block; i.e., $d=f$. Here

$$x_{cd} = \frac{(k-1)}{mk(k-2)}\{(k-1)C_{cd}+C_{ed}\} \quad \ldots . \quad (7.5)$$

$$x_{ed} = \frac{(k-1)}{mk(k-2)}\{(k-1)C_{ed}+C_{cd}\}$$

6. There are p treatments (not c or e) common to blocks d and f. It is possible that $p=0$. In addition, x_{cd} and x_{ef} occur in the same treatment; i.e., $c=e$. Then

$$x_{cd} = \frac{1}{m^2-(t-1)^2[(k-1)^2+p]^2}\{mC_{cd}+(t-1)[(k-1)^2+p]C_{cf}\} \quad \ldots . \quad (7.6)$$

$$x_{cf} = \frac{1}{m^2-(t-1)^2[(k-1)^2+p]^2}\{mC_{cf}+(t-1)[(k-1)^2+p]C_{cd}\}$$

Example

The example presented in this section utilises data from an experiment in which alfalfa was subjected to nine top dressings of fertiliser. The observations represent yields of cured hay in pounds per fifty square feet. The design was a balanced incomplete block with $r=4$, $t=9$, $k=3$, $b=12$, and $\lambda=1$; λ is the number of times a particular pair of treatments appear together in a block. There were two missing observations: x_{34}, in treatment 3 and block 4, and x_{37}, in treatment 3 and block 7. These two cells are 'boxed in' in Table II.

To analyse a balanced incomplete block design with missing observations, one first obtains the original block totals and the original grand

total, shown in Table II. Next the original quantities T_i and $B_{.i}$ are obtained at the same time on a calculator by splitting the keyboard and as one is adding the observations to obtain T_i on the right-hand side of the keyboard, the B_j's can be summed on the left-hand side to obtain the $B_{.i}$'s. Then the original Q-values are obtained from (2.2), e.g.,

$$Q_3' = 3(8{\cdot}69) - 64{\cdot}93 = -38{\cdot}86$$

The original values of T_i, $B_{.i}$, and Q_i are given in Table III. (Some of the original values will of course be T_i', $B_{.i}'$, and Q_i'.) For a computational check it should be noticed that the total of the $B_{.i}$'s is equal to k times the grand total and that the sum of the Q_i's should be zero.

Next the configuration of the missing values should be determined. In this problem the two missing values occur in the same treatment, but

TABLE II

Original data and block totals

Treatment Nos. and Observations			Original Totals	Augmented Totals
4 8·21	5 6·78	6 6·73	B_1 21·72	B_1 21·72
7 6·01	8 9·13	9 6·89	B_2 22·03	B_2 22·03
1 7·39	2 6·04	3 3·87	B_3 17·30	B_3 17·30
3 [4·71]	6 6·40	9 6·78	B_4' 13·18	B_4 17·89
2 7·75	5 7·23	8 7·57	B_5 22·55	B_5 22·55
1 7·66	4 7·94	7 6·11	B_6 21·71	B_6 21·71
3 [4·89]	4 8·36	8 8·34	B_7' 16·70	B_7 21·59
1 7·67	5 6·07	9 6·51	B_8 20·25	B_8 20·25
2 7·57	6 6·38	7 5·96	B_9 19·91	B_9 19·91
1 7·81	6 7·35	8 8·56	B_{10} 23·72	B_{10} 23·72
3 4·82	5 6·82	7 6·11	B_{11} 17·75	B_{11} 17·75
2 7·03	4 8·31	9 6·11	B_{12} 21·45	B_{12} 21·45
			G' 238·27	G 247·87

TABLE III

Treatment totals, $B_{.i}$'s and Q_i's

Treatment No.	T_i		$B_{.i}$		Q_i	
	Original	Augmented	Original	Augmented	Original	Augmented
1	30·53	30·53	82·98	82·98	8·61	8·61
2	28·39	28·39	81·21	81·21	3·96	3·96
3	8·69	18·29	64·93	74·53	−38·86	−19·66
4	32·82	32·82	81·58	86·47	16·88	11·99
5	26·90	26·90	82·27	82·27	−1·57	−1·57
6	26·86	26·86	78·53	83·24	2·05	−2·66
7	24·19	24·19	81·40	81·40	−8·83	−8·83
8	33·60	33·60	85·00	89·89	15·80	10·91
9	26·29	26·29	76·91	81·62	1·96	−2·75
Total	238·27	247·87	714·81	743·61	0·00	0·00

there are no other treatments replicated in the two involved blocks. Hence, one would apply the formulae (7.6), with $p=0$. To apply these formulae we must first compute from (3.2) and (2.5)

$$m = (3-1)[(4)(9)(3) - (4)(9) - (9)(3) + 3] = 96$$

$$\begin{aligned} C_{34} &= rt(k-1)B'_4 + (t-1)(k-1)Q'_3 - (t-1)[Q'_{6(3)} + Q'_{9(3)}] \\ &= (4)(9)(2)(13{\cdot}18) + (8)(2)(-38{\cdot}86) - (8)(2{\cdot}05 + 1{\cdot}96) \\ &= 295{\cdot}12 \end{aligned}$$

and

$$\begin{aligned} C_{37} &= rt(k-1)B'_7 + (t-1)(k-1)Q'_3 - (t-1)[Q'_{4(3)} + Q'_{8(3)}] \\ &= (4)(9)(2)(16{\cdot}70) + (8)(2)(-38{\cdot}86) - (8)(16{\cdot}88 + 15{\cdot}80) \\ &= 319{\cdot}20 \end{aligned}$$

Then the missing value estimates are obtained from (7.6):

$$x_{34} = \frac{mC_{34} - (t-1)(k-1)^2 C_{37}}{m^2 - (t-1)^2(k-1)^4}$$

$$= \frac{(96)(295{\cdot}12) + (8)(4)(319{\cdot}20)}{96^2 - (8^2)(2^4)} = 4{\cdot}71$$

$$x_{37} = \frac{mC_{37} + (t-1)(k-1)^2 C_{34}}{m^2 - (t-1)^2(k-1)^4}$$

$$= \frac{(96)(319{\cdot}20) + (8)(4)(295{\cdot}12)}{96^2 - (8^2)(2^4)} = 4{\cdot}89$$

The values of x_{34} and x_{37} are entered in Table II and then the augmented quantities B_j, T_i, $B_{.i}$, Q_i, and G are obtained and used to compute the sums of squares for the approximate analysis.

Since the insertion of missing values results in a biased treatment sum of squares, the usual procedure of computing the block and total sums of squares is modified so that the exact test of treatment effects can be easily obtained. The block sum of squares is obtained as

$$\text{Block S.S. (unadjusted)} = \frac{1}{k}\Sigma B_j^2 + \frac{1}{k}\Sigma[B_j' + \Sigma_j x_{ij}]^2 - \frac{G^2}{N}$$

$$= \frac{4380{\cdot}4699}{3} + \frac{786{\cdot}1802}{3} - 1706{\cdot}6538$$

$$= 15{\cdot}5629$$

where the first summation refers to blocks with no missing values and the second to the augmented totals of blocks with missing values. The total sum of squares is obtained as

$$\text{Total S.S.} = \Sigma\Sigma y_{ij}^2 + \Sigma\Sigma x_{ij}^2 - \frac{G^2}{N}$$

$$= 1709{\cdot}5133 + 46{\cdot}0962 - 1706{\cdot}6538$$

$$= 48{\cdot}9557$$

where the first summation refers to observed values and the second to estimates of missing values.

The treatment sum of squares is obtained in the usual manner:

$$\text{Treatment S.S. (adjusted)} = \frac{(t-1)}{rtk(k-1)}\ \Sigma Q_i^2$$

$$= 834{\cdot}1894/27$$

$$= 30{\cdot}8959$$

and the error sum of squares is obtained by subtraction. The above sums of squares are then set up as in Table IV, care being taken to subtract, from both the error and total degrees of freedom, one for each missing value estimated.

TABLE IV

Approximate analysis of variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Blocks (Unadjusted)	11	15·5629	1·4148	
Treatments (Adjusted)	8	30·8959	3·8620	21·65
Error	14	2·4969	0·1784	
Total	33	48·9557		

The approximate test indicates significance at the 0·1 % level, the appropriate tabular value of F being 6·80; since this significance may be due to the bias, it is necessary to perform the exact analysis. If the approximate test were not significant one could stop here.

To obtain the exact analysis, the design is considered as a one-way classification ignoring treatments and the appropriate block and total

sum of squares are calculated using only the original data and making use of the quantities already computed.

$$\text{Block S.S.} = \frac{4380{\cdot}4699}{3} + \frac{[(13{\cdot}18)^2 + (16{\cdot}70)^2]}{2} - \frac{(238{\cdot}27)^2}{34}$$

$$= 16{\cdot}6757$$

$$\text{Total S.S.} = 1709{\cdot}5133 - \frac{(238{\cdot}27)^2}{34}$$

$$= 39{\cdot}7312$$

The exact treatment sum of squares is then obtained by subtracting the error sum of squares of Table IV and the block sum of squares for the one-way classification from the total sum of squares for the one-way classification. These sums of squares are shown in Table V.

TABLE V

Exact analysis of variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F*
Blocks	11	16·6757	1·5160	
Treatments (Adjusted)	8	20·5586	2·5698	14·40
Error	14	2·4969	0·1784	
Total	33	39·7312		

Although the calculated *F*-value was reduced, the exact test still indicates significance at the 0·1 % level.

REFERENCES

[1] Cornish, E. A. (1940). 'The estimation of missing values in incomplete randomized block experiments', *Annals of Eugenics*, **10**, 112.

[2] Glenn, W. A. and Kramer, C. Y. (1958). 'Analysis of variance of a randomised block design with missing observations', *Applied Statistics*, **7**, 173.

[3] Wilkinson, G. N. (1958). 'Estimation of missing values for the analysis of incomplete data', *Biometrics*, **14**, 257.

MEETINGS OF SECTIONS AND LOCAL GROUPS OF THE ROYAL STATISTICAL SOCIETY

INDUSTRIAL APPLICATIONS SECTION

5th February 1960. The Analysis of Censored Data: J. Taylor

1st April 1960. Some Aspects of Specifications and Tolerances: R. H. S. Lesser

6th May 1960. Industrial Applications of Time Series: G. A. Barnard

GENERAL APPLICATIONS SECTION

10th February 1960. Some Problems in the Coding of Rare Events: J. P. Martin

10th March 1960. The Rise and Decline of Spinsterhood: J. Hajnal

13th April 1960. Why do People Move House?: D. V. Donnison

11th May 1960. The Relationship between Unemployment and the Rate of Change of Money Wage Rates in the U.K. between 1862 and 1957: R. G. Lipsey

MEDICAL SECTION

26th January 1960. Mechanisms of Carcinogenesis—Epidemiological Evidence: R. S. Doll

23rd February 1960. A Statistical Approach to the Diagnosis and Assessment of Therapeutic Response in Thyroid Disease: J. Crooks

29th March 1960. Some Statistical Problems in Human Genetics: C. A. B. Smith

26th April 1960. Susceptibility to Acute Respiratory Diseases in Children: E. G. Knox

27th May 1960. A Symposium on Related Variables in Medical Statistics —Multiple Regression in Hospital Costing: N. T. J. Bailey; Discriminant Functions in the Diagnosis of Lung Cancer: T. H. Hollingsworth; Factor Analysis in Clinical Psychology: A. E. Maxwell

BIRMINGHAM AND DISTRICT GROUP

6th January 1960. Selling Statistics to Industry: H. C. Hamaker

4th February 1960. Sequential Inspection: K. W. Kemp

2nd March 1960. Operational Research in Industry: K. D. Tocher

6th April 1960. Routine Analysis of Data on a Computer: D. H. Rees

BRISTOL GROUP

14th January 1960. The Role of Statistics in Operational Research: K. D. Tocher

28th January 1960. Statistical Problems in General Practice: R. J. F. H. Pinsent (Joint Meeting with the South West England Faculty of the College of General Practitioners and the West of England Branch of the Society of Medical Officers of Health)

18th February 1960. On Estimating Mortality in Wildfowl Populations: H. J. Boyd

17th March 1960. Short-Cut Approximate Methods in Statistical Analysis: P. G. Moore

EDINBURGH GROUP

19th January 1960. Collecting Farming Statistics in Scotland: P. M. Scola
16th February 1960. Industrial Applications of Time Series: G. A. Barnard
8th March 1960. A Case History in Operational Research: K. D. Tocher
6th April 1960. Peri-natal Mortality: A. F. Purser
4th May 1960. The Analysis of Censored Data: M. R. Sampford

GLASGOW GROUP

26th January 1960. Statistical Methods in Brewing Research: I. A. Preece
18th February 1960. Stock Control: J. Battersby
29th March 1960. A Traffic Survey in the Central Area of Glasgow: T. B. McKenna

LEICESTER GROUP

28th January 1960. The Analysis of Censored Data: J. Taylor

MERSEYSIDE GROUP

12th January 1960. Life Tests with Two Sources of Failure: R. L. Plackett
9th February 1960. Queues and Machine Interference: H. F. Downton
8th March 1960. Statistical Tests of the Accuracy of Weighing: D. H. Ward
12th April 1960. The Effort of Inspection: S. R. Broadbent

NORTH-EASTERN GROUP

13th January 1960. Non-Parametric Methods: P. A. Wallington
17th February 1960. Industrial Applications of Time Series: G. A. Barnard
9th March 1960. Asking Questions: N. J. Squirrell

SHEFFIELD GROUP

28th January 1960. Evolutionary Operation: G. A. Coutie
25th February 1960. The Analysis of Failure Data: D. R. Cox
31st March 1960. Bayesian Statistics: D. V. Lindley
28th April 1960. Some Programmes on the Analysis of Experiments: K. D. Tocher

SOUTH WALES GROUP

7th January 1960. Statistical Assessment of Attitudes towards Management: R. Revons
4th February 1960. Uncertainty and Probability: G. L. S. Shackle
8th March 1960. How to Cure the Common Cold: P. D. Oldham

BOOK REVIEWS

S. K. Ekambaram: *The Statistical Basis of Quality Control Charts.* London: Asia Publishing House, 1960. Pp. x + 96. 16s. 6d.

The dust-cover of this book states that it offers 'a simple but rigorous presentation of the basis of the Statistical Quality Control'; the preface states that it 'is written for the busy executive in the higher levels of Management who desires to have the general picture of the technicalities of quality control in industry, but has no time to go into the details of the quality control work'. My impression of the work is that in places it is anything but simple, that in places it is anything but rigorous, and that the busy executive would very soon decide that he was really too busy to waste his time on it.

As an example of lack of simplicity the beginning of Chapter V is worth quoting:

'The statistical universe of production would be homogeneous (i.e. only one universe) when there is a single standard of quality. In the case of a measurable quality, we expect the quality measure to cluster near a value which may be standard level of quality, and also the variation about that level to be something which can be estimated. When the quality is steady and standard, it is the entire aggregate of production that can help us to evaluate the standard. The pattern of variation of the quality (measure) in the aggregate describes the quality best.' What will the busy executive make of that?

Another example of lack of simplicity is (page 12) a list of measures of dispersion: 'inter-quartile range, the coefficient of variation, the average deviation and the standard deviation' without any explanation of what these measures are, except for the standard deviation for which the formula is given, but no attempt is made really to explain even this. It would have helped if the picture of the normal curve had stated that the scale was in units of the standard deviation, but this scale is just marked 1, 2, 3 without any units.

As an example of lack of rigour, the factors A_2, D_3, D_4 and d_2 (with which to multiply the average sample range for various purposes) are introduced (page 35) without any explanation other than that they have been found useful, and that values for various sample sizes can be found in technical books on statistical quality control.

Comprehension is also not helped by the oddities of language which, while most of them are trivial in themselves, tend to have a cumulative effect, e.g. 'The range is used in small samples, say, up to a size of six or less'; '$\mu_4/\mu_2{}^2$ is defined as β_2' when what is meant is 'β_2 is defined as $\mu_4/\mu_2{}^2$'; 'satisfactory ones are called as "non-defectives" '; 'situations when these need not both be $= \frac{1}{2}$', and so on.

The printing and proof reading are also far from perfect. On page 69 the multiplier, 3, jumps in and out of the square-root sign with great abandon from one line to the next. But the biggest muddle occurs in Chapters XI and XII, on the binomial distribution. Here the binomial coefficients,

${}_nC_r$, appear on pages 60, 61, and 62 (except for two cases which are correctly printed) as nC_r. Thus we find, for example,

$$(\tfrac{1}{2}+\tfrac{1}{2})^3 = (\tfrac{1}{2})^3 + 3C_1(\tfrac{1}{2})^2(\tfrac{1}{2}) + 3C_2(\tfrac{1}{2})(\tfrac{1}{2})^2 + 3C_3(\tfrac{1}{2})^3$$
$$= \tfrac{1}{8} + 3\tfrac{1}{8} + 3\tfrac{1}{8} + \tfrac{1}{8}.$$

On page 63, the n and the r get into the same line (most of the time), but it is the wrong line. They are now on a level with the p's and q's, the C going up into the index line, thus: ${}^nCn-r\ q^{n-r}p$ and this makes the algebra very hard to read as there are no brackets around the $(n-r)$ terms, and the minuses of these subscripts come in the same line as the pluses between one term and the next. On pages 65 and 66 the situation is further confused.

In addition, the book is definitely misleading in places. If I had read it without previous knowledge, I should have gained the impression that industrial experience showed the normal law to be true for all directly measurable qualities (page 21), that rational sub-groups were usually obtained by repeat measurements of the same article (page 35), and that a control chart 'in control' proved that only chance causes were operating (page 39).

The definition of a random sample (page 30) is incomplete since it only requires 'that every unit in the aggregate has an equal chance of being chosen in the sample' and does not mention that each possible combination of units should have an equal chance.

Finally, two direct mistakes: (i) it is stated (page 22) that for a normal curve 'f(x) has the same value at $+(x-\mu)$ and at $-(x-\mu)$'; (ii) it is stated (page 23) that at the points of inflexion at $\mu+\sigma$ and $\mu-\sigma$ 'the curve changes the sign of its "slope" '.

I. D. HILL

D. S. Falconer: *Introduction to Quantitative Genetics*. Edinburgh and London: Oliver and Boyd, 1960. Pp. ix + 365. 35s.

This book provides an excellent introduction to the subject of quantitative genetics without suffering from oversimplification. The author provides sufficient glimpses of controversial topics to whet the appetite without being drawn too far into issues that are still obscure or still the subject of much debate. Where necessary he is careful to inform the reader of matters which are his personal opinion while at the same time giving references to the work of others. Certain knowledge is assumed on the part of the reader and this is explained in the preface. The book can be read with profit by anyone who is at ease with the methods of analysis of variance, regression, and correlation, as a good practical knowledge of these methods is assumed. Anyone to whom these matters are still the subject of some mystery would do well to delay reading this book until the methods have been mastered. It is hoped that readers of this review will not take this to mean that the book is highly mathematical, because it is not, but familiarity with the techniques mentioned will ensure that the genetic ideas are more readily assimilated. Only an elementary knowledge of Mendelian genetics is assumed and for this and the above reasons I think that the book has greater appeal to the statistician with little knowledge of genetics than to a geneticist with little knowledge of practical statistics.

Throughout the book there are a number of worked examples taken from actual experimental data which are used to demonstrate the theory as it is expounded. The reader would do well to check these as he reads, as there

are a number of misprints scattered throughout the volume which might well cause trouble if not detected. It is to be hoped that these will be corrected when the work comes to a second edition.

The first chapter deals with the well known Hardy-Weinberg equilibrium for large populations. The author then proceeds to demonstrate how this can be disturbed in turn by migration, mutation, and selection, and finally obtains results taking into account all of these influences at once. This method of considering individual factors one at a time and then compounding them all into the final result makes the book very clear and exceedingly readable. There are three chapters on the behaviour of gene frequencies in small populations followed by a further three chapters introducing metrical characters. By this time the reader is prepared for a study of genetic covariance to explain the resemblance between relatives, and a further four chapters on selection experiments and the results obtained. There are then three chapters on the effects of inbreeding and crossbreeding followed by one each on scaling problems, threshold characters, and correlated characters. The book concludes with a discussion of metric characters and variation under conditions of natural selection.

This book does not (as the author explains) provide an exhaustive list of references and should not be consulted for this purpose. However there are many references in the text to recent work upon the matters being discussed. The presentation is such that the reader is powerfully persuaded that this is a subject that is very much alive, and that although the body of accumulated knowledge is large there remain many exciting discoveries to be made.

It is a most readable and informative book.

A. M. O. VEALE

M. H. Quenouille: *Rapid Statistical Calculations: A Collection of Distribution-free and Easy Methods of Estimation and Testing.* London: Charles Griffin, 1959. Pp. xv + 81. 10s.

Statisticians are sometimes called on to make quick estimates or to give quick answers to questions about the significance of experimental results. If, when confronted with such requests, the statistician has to waste much time looking up a method and finding out whether it is appropriate, the quick method ceases to be quick, as the author points out. Hence he has collected a number of rapid methods, and has presented them in a brief straightforward fashion in this pocket manual, which lives up to this description, being small and slim. Yet in spite of the small size, clarity of presentation has been achieved; each method and an example of its use are given on a pair of facing pages.

It would be tedious to list all the tests. The main classification is into (*a*) methods involving mean level or scatter of groups of observations; (*b*) methods comparing frequencies or proportions; and (*c*) methods involving association. An appendix gives abbreviated versions of a number of common statistical tables.

It may be remarked here that the sub-title refers to part of the collection rather than to the whole. In each section the most efficient test is first given; such a test will in some cases not be distribution-free. However, the inclusion of the efficient tests is certainly a good thing; in many situations the statistician is able to carry out pencil and paper calculations quite rapidly with simple data; and the most efficient method may well be the proper one to use, even though it is not the quickest.

After the most efficient tests come those of lower efficiency. The use of these may be justified on several grounds. They may be distribution-free and so applicable in situations where the corresponding efficient test is not strictly valid. They may be so much quicker than the efficient method that the use of the latter is not worth while. (In particular, counting methods are quick and are useful when it is difficult to carry out even the simplest calculations.) A quick method may be a useful guide to the desirability of performing a full analysis, or a check on the latter when it has been performed.

An indication of the efficiency of most of the tests has been given, and in addition they have been classified as quick, very quick, faster than usual, and distribution-free. No claim for completeness is made; the author has given the tests he believes to be most useful.

Statisticians vary in their reaction to quick-and-easy tests; some revel in them, and others feel vaguely uneasy when using them. I shall not attempt to take sides but am content to commend this book as a useful compilation which admirably achieves what it sets out to do. D. G. BEECH

J. M. Sengupta and Nikhilesh Bhattacharya: *Tables of Random Normal Deviates*. London: Asia Publishing House; Calcutta: Statistical Publishing Society, 1960. Pp. xix + 26. 35s.

This series of random normal deviates is issued under the auspices of the Indian Statistical Institute. It consists of 10 400 deviates arranged in 26 plates, and the values were derived from L. H. C. Tippett's series of random numbers. Hence they are less extensive than Wold's series or the RAND Corporation series, which list 25 000 and 100 000 deviates respectively, but they form an additional source of supply.

As listed they are random normal deviates with zero mean and standard deviation 1000, but conversion to unit standard deviation is, of course, simple. However, instead of each deviate being prefixed with a plus or a minus sign, the digits 1 and 2 are used as prefixes, 1 indicating plus and 2 indicating minus. It seems a pity that the proper signs were not used, especially with such a relatively expensive book. Any device which increases the risk of error is to be regretted.

Various tests carried out on the tables are described and the results are given. They are described as 'satisfactory' or 'fairly satisfactory'. In the latter category it appears that the mean of the 10 400 values, converted to unit standard deviation, is $-0{\cdot}02238$; the difference from zero is significant at about the 2% level (two-tailed test). Personally I should feel uneasy about this when using the tables. D. G. BEECH

INDEX

Acton, F. S., 130

Bailey, D., 51
Bailey, N. T. J., 64
Baird, H. R., 189
Batchelor, J. H., 66
Beech, D. G., 64, 68, 132, 204
Bhattacharya, N., 204
Birnbaum, A., 152
Bliss, C. I., 8
Booker, H. S., 63
Bryson, M. R., 178
Buckland, W. R., 63
Burdess, T. K., 113

Calder, A. B., 170
classification based on Bayes's formula, 152
consumer panel, potential biases, 20
Corlett, E. N., 92
Coutie, G. A., 61

Dodge, H. F., 67
Duckworth, W. E., 133

'Economics for Pleasure', review, 62
economics of new roads, 113
Ehrenberg, A. S. C., 20
Ekambaram, S. K., 201

Falconer, D. S., 202
fractional factorial designs, 61
frequency counts, comparison of two, 37
Freund, J. E., 67

geology, statistics in, 82
Glass, Suzanne, 43
Gregory, G., 92
Gridgeman, N. T., 103

handwheel, consistency of setting, 92
Hill, I. D., 69, 202
Holland, D. A., 1

Ilersic, A. R., 68
instrumental analysis, 170
'International Journal of Abstracts—Statistical Theory and Method', review, 63
interviewer variability in a budget survey, 122
inventory using sampling, 178

Kemsley, W. F. F., 122
Kramer, C. Y., 43, 189
Krumbein, W. C., 82
Kupperman, M., 37

linear programming in agricultural economics, 28

Maxwell, A. E., 152
missing observations:
in balanced incomplete blocks, 189
in Latin square design, 43
'Modern Business Statistics', review, 67
Muir, A., 66
multivariate methods in botany, 1

'Operations Research: an Annotated Bibliography', review, 66

Pearce, S. C., 1
preference tests, 8
'Public and the Programmes, The', note, 132

'Quantitative Genetics, Introduction to', review, 202
Quenouille, M. H., 203

'Rapid Statistical Calculations', review, 203
'Regression Analysis', review, 130
Ridley, T. M., 67
Romig, H. G., 67

sampling inspection, economic incentive, 69
'Sampling Inspection Tables, Single and Double Sampling', review, 67
Sengupta, J. M., 204
sensory threshold, 13
Shackle, G. L. S., 62
Smith, C. A. B., 66
'Statistical Basis of Quality Control Charts, The', review, 201
'Statistical Methods in Biology', review, 64
'Statistics', note, 68
stock control, 133
'Straight-Line Data, Analysis of', review, 130

'Tables of Random Normal Deviates', review, 204
taste testing, 103

unemployment statistics, British, 51

Veale, A. M. O., 203
Vincent, Shirley E., 28

Williams, E. J., 130
Williams, F. J., 67

Printed in Great Britain by
ROBERT CUNNINGHAM AND SONS LTD.
Alva